UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

MESTRADO E DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO

MÁRCIO ANDRÉ SOUTO CORREIA

AVALIAÇÃO DE FEATURES DE LOCALIZAÇÃO PARA AUTENTICAÇÃO TRANSPARENTE E CONTÍNUA: PROCESSO E ESTUDO DE CASO

FORTALEZA

2016

MÁRCIO ANDRÉ SOUTO CORREIA

AVALIAÇÃO DE FEATURES DE LOCALIZAÇÃO PARA AUTENTICAÇÃO TRANSPARENTE E CONTÍNUA: PROCESSO E ESTUDO DE CASO

Dissertação de Mestrado submetida à Coordenação do Programa de Pós-Graduação em Ciência da Computação (MDCC) da Universidade Federal do Ceará, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.
Orientadora: Profa. Rossana M. C. Andrade, PhD.

FORTALEZA

2016

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C848a Correia, Márcio André Souto.
Avaliação de Features de Localização para Autenticação Transparente e Contínua : Processo e Estudo de Caso / Márcio André Souto Correia. – 2016.
101 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2016.
Orientação: Profa. Dra. Rossana Maria de Castro Andrade.

1. Autenticação. 2. Biometria. 3. Avaliação de Desempenho. 4. Aprendizado de Máquina. 5. Computação Móvel. I. Título.

CDD 005

MÁRCIO ANDRÉ SOUTO CORREIA

AVALIAÇÃO DE FEATURES DE LOCALIZAÇÃO PARA AUTENTICAÇÃO TRANSPARENTE E CONTÍNUA: PROCESSO E ESTUDO DE CASO

Dissertação de Mestrado submetida à Coordenação do Programa de Pós-Graduação em Ciência da Computação (MDCC) da Universidade Federal do Ceará, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Aprovada em: 30 de Novembro de 2016

BANCA EXAMINADORA

Profa. Rossana M. C. Andrade, PhD. (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Paulo Lício Geus, Dr.
Universidade Estadual de Campinas (UNICAMP)

Prof. André Luiz Moura dos Santos, Dr.
Universidade Estadual do Ceará (UECE)

Prof. José Antônio Fernandes de Macêdo, Dr.
Universidade Federal do Ceará (UFC)

À minha família, especialmente Renata Teixeira (esposa) e Melina Correia (filha). Sem vocês nada disso seria possível.

AGRADECIMENTOS

Aos meus familiares, especialmente minha esposa Renata e filha Melina, mas também meus pais (Edmar e Fátima), meus sogros (Heldon e Célia) e meus irmãos (Marcos e Michell), pelo amor, incentivo e suporte que foram fundamentais para superação dos mais importantes desafios deste projeto.

À Profª. Rossana Andrade, pela orientação, oportunidades, tempo dedicado, confiança e os valiosos ensinamentos ao longo de toda esta jornada.

Ao Prof. André Santos, pelo incentivo desde a graduação na UECE, quando tive a oportunidade de ser orientado por ele durante a produção da minha monografia de conclusão do curso.

Ao Prof. José Macedo, pela confiança depositada em mim quando esteve a frente da Secretaria de Tecnologia da Informação (STI) da UFC e, sob a sua direção, tive a oportunidade de participar de vários projetos importantes para a Universidade e também iniciar o mestrado.

Ao Prof. Paulo Lício, pela disponibilidade em avaliar e contribuir com a melhoria desta dissertação.

À equipe da STI, principalmente o Prof. Joaquim Bento (Diretor Executivo), Prof. José Ramos (Diretor Adjunto), Amarildo Rolim (Diretor da Divisão de Segurança da Informação) e a equipe da Assessoria de Governança (Esrom Bomfim, Evelyne Avelino, Lucas Magalhães, Matheus Costa), pelo apoio e incentivo que também foram fundamentais neste projeto.

Aos professores, alunos e colaboradores do GREat e do MDCC, principalmente: aos professores Carlos Fisch, Fernando Trinta, José Neuman, Victor Campos e Windson Viana, pelo conhecimento repassado nas disciplinas ministradas; aos alunos de doutorado Rainara Maia, Rayner Gomes e Juliana Oliveira, pela parceria nas disciplinas; aos professores Miguel Franklin e Danielo Gomes, pela avaliação deste trabalho durante as etapas de proposta e qualificação e as sugestões de melhoria; aos professores Márcio Maia, Paulo Armando e Lincoln Rocha e aos alunos de doutorado Carlos André Batista e Danilo Reis, pelas conversas sobre este trabalho e as sugestões de melhoria; e às colaboradoras Darilu Martins e Janaína Bezerra, pelo apoio e acolhimento ao longo desses anos.

Aos amigos pessoais pelo incentivo, principalmente aos parceiros nessa empreitada Pablo Ximenes e Wagner Marques, pelas incontáveis horas de conversas na tentativa de superar os desafios encontrados na condução das nossas respectivas pesquisas.

“Quem caminha sozinho pode até chegar mais rápido, mas aquele que vai acompanhado, com certeza vai mais longe.”

(Clarice Lispector)

RESUMO

Com um número cada vez maior de dispositivos ao redor dos usuários e, ao mesmo tempo, aplicações que demandam interações mais rápidas e frequentes com esses equipamentos, novos mecanismos buscam reduzir o tempo gasto pelos usuários com autenticação e aumentar o nível de segurança relacionado com a verificação de identidade. Nesse sentido, existem várias propostas na literatura com mecanismos de autenticação transparente e contínua que combinam dados biométricos obtidos a partir de ações que os usuários já realizam enquanto usam os dispositivos móveis (e.g. localização, toque na tela, digitação, marcha, voz, entre outras). Na revisão da literatura realizada neste trabalho, foram encontradas nove propostas que usam a localização *outdoor* do usuário combinada com outros tipos de *features* biométricas como entrada para o mecanismo de autenticação proposto. Essas propostas têm em comum não só o uso da localização *outdoor*, mas também o fato de não conseguirem avaliar de maneira adequada cada conjunto de *features* individualmente. Sendo assim, este trabalho tem como objetivo propor um processo de avaliação de *features* biométricas, adaptando diretrizes de aprendizado de máquina, para realização de experimentos com base em uma metodologia estatística. Esse processo de avaliação é importante para a completa compreensão do funcionamento do mecanismo proposto, o que permite a identificação e o reuso das técnicas de extração de *features* que oferecem melhor desempenho. Além disso, um estudo de caso usando o processo é realizado para a avaliação e a comparação das *features* de localização *outdoor* identificadas na literatura. Nessa avaliação foram realizados experimentos com três algoritmos de classificação (C4.5, SVM e Naive Bayes) disponíveis no ambiente de aprendizado de máquina WEKA e quatro conjuntos de dados, sendo dois deles públicos (Geolife e MIT Reality). Foram também coletadas doze medidas, sendo nove delas de eficácia e três de eficiência. A análise dos resultados dos experimentos mostrou variações significativas na acurácia, uso de CPU e memória, considerando todos os cenários avaliados. Com esses resultados, este trabalho fornece evidências sobre a viabilidade do processo proposto, produzindo resultados para guiar a escolha de *features* de localização *outdoor* e algoritmos de aprendizado que oferecem melhor desempenho para a construção de mecanismos de autenticação transparente e contínua.

Palavras-chave: Autenticação; Biometria; Avaliação de Desempenho; Aprendizado de Máquina; Computação Móvel.

ABSTRACT

Given today's growing number of devices around users, and, at the same time, their faster and frequent interactions with these devices, new security mechanisms have emerged aiming at reducing the time spent by users with authentication as well as raising the security level related to identity verification. In this sense, there are several proposals in the literature with transparent and continuous authentication mechanisms that combine biometric data retrieved from actions that users already do while using mobile devices (e.g. location, screen touch, keystroke, gait, voice, among others). In the literature review performed in this work were found nine proposals that use outdoor location and merge other kinds of biometric features as input to their proposed authentication mechanism. These proposals have in common not only the use of outdoor location but they also fail to evaluate properly each biometric features set individually. Therefore, this work provides a new process for evaluation of biometric features by adapting guidelines of machine learning to perform experiments based on a statistical methodology. This is important to know how the mechanism works, which allows the identification and reuse of features extraction techniques that provide the best performance. Moreover, this process is also used in this work to evaluate and compare the outdoor location features identified in literature. For this evaluation, experiments were conducted with three classification algorithms (C4.5, SVM, and Naive Bayes) available in the WEKA machine learning environment and four datasets, two of which are public (Geolife and MIT Reality). Besides that, twelve measures were collected, being nine efficacy and three efficiency measures. In the analysis of the experimental results, significant variations were found in accuracy, CPU time, and memory regarding all evaluated scenarios. With these results, this work provides evidence of the viability of the proposed process and guides the choice of outdoor location features and learning algorithms that provide better performance for constructing transparent and continuous authentication mechanisms.

Keywords: Authentication; Biometrics; Performance Evaluation; Machine Learning; Mobile Computing.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AAA	*Authentication, Authorization, and Accountability*
ACM	*Association for Computing Machinery*
ARFF	*Attribute-Relation File Format*
AUC	*Area Under Curve*
BEAT	*Biometrics Evaluation and Testing*
BPMN	*Business Process Model and Notation*
CPU	*Central Processing Unit*
DNA	*Deoxyribonucleic Acid*
DT	*Decision Tree*
EER	*Equal Error Rate*
FAR	*False Acceptance Rate*
FMR	*False Match Rate*
FN	*False Negative*
FNMR	*False Non-Match Rate*
FP	*False Positive*
FRR	*False Rejection Rate*
FTA	*Failure To Acquire*
FTE	*Failure To Enroll*
GPS	*Global Positioning System*
IBG	*International Biometrics Group*
IEC	*International Electrotechnical Commission*
IEEE	*Institute of Electrical and Electronics Engineers*
ISO	*International Organization for Standardization*
MIT	*Massachusetts Institute of Technology*
NB	*Naive Bayes*
PIN	*Personal Identification Number*

RAM	*Random Access Memory*
RBF	*Radial Basis Function*
ROC	*Receiver Operating Characteristic*
SVM	*Support Vector Machine*
TAR	*True Acceptance Rate*
TN	*True Negative*
TP	*True Positive*
TRR	*True Rejection Rate*
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Este trabalho apresenta uma proposta de processo e um estudo de caso focados na avaliação de *features* biométricas para autenticação transparente e contínua em dispositivos móveis usando aprendizado de máquina. As próximas seções deste capítulo estão estruturadas da seguinte forma: a seção 1.1 apresenta a contextualização; em seguida, a seção 1.2 apresenta a motivação e descreve o problema; a seção 1.3 estabelece os objetivos; a seção 1.4 define os resultados esperados; a seção 1.5 apresenta a metodologia utilizada; e por fim, a seção 1.6 apresenta a organização dos demais capítulos deste trabalho.

1.1 Contextualização

Renaud (2005) recomenda que as soluções de autenticação sejam projetadas para equilibrar segurança e usabilidade. Tradicionalmente, busca-se garantir o maior nível possível de complexidade para as senhas, tendo como limite, por exemplo, a capacidade de memorização do usuário. Nesse sentido, algumas variáveis importantes incluem o comprimento da senha e o conjunto de símbolos do alfabeto usado para a sua composição. Assim, senhas escolhidas pelo usuário com 4 caracteres numéricos podem garantir um nível adequado de usabilidade, pois são fáceis para um usuário memorizar. No entanto, podem não garantir um nível adequado de segurança, pois também são fáceis para um adversário adivinhar. Por outro lado, senhas geradas aleatoriamente com mais de 12 caracteres alfanuméricos, incluindo obrigatoriamente letras maiúsculas, minúsculas, números e caracteres especiais, podem melhorar a segurança. Entretanto, podem prejudicar excessivamente a usabilidade, pois são difíceis para um usuário memorizar e digitar.

Alguns pesquisadores da área, como Jakobsson *et al.* (2009) e Crawford (2014), acreditam que os dispositivos móveis dificultam o uso de senha por causa de limitações impostas por sua interface de entrada. O estudo realizado por Melicher *et al.* (2016) traz elementos que apoiam essa hipótese. Eles compararam a autenticação baseada em senhas entre os dispositivos móveis (*i.e. smartphone* e *tablet*) e os dispositivos tradicionais (*i.e. desktop* e *laptop*), revelando que os dispositivos móveis impõem restrições significativas para a segurança e a usabilidade das senhas. Em termos de segurança, os resultados desse estudo revelaram que as senhas criadas em dispositivos móveis são 32% mais fracas. Além disso, em termos de usabilidade, esse mesmo estudo indicou que os usuários de dispositivos móveis gastam 20% mais tempo na digitação de

senhas e falham 100% mais por erros de digitação.

O problema é ainda mais grave se considerarmos uma pesquisa online promovida pelas empresas Lookout e Sprint (2014) que buscou identificar os hábitos de segurança dos usuários de dispositivos móveis. Essa pesquisa concluiu que 56% dos proprietários de *smartphones* não configuram senha em seus dispositivos. Já em 2009, Jakobsson *et al.* (2009) também pesquisaram os hábitos de segurança dos usuários de dispositivos móveis e descobriram que eles acham mais irritante digitar senhas nesses dispositivos do que a ausência de cobertura de rede, a tela de tamanho pequeno ou a qualidade ruim da voz durante as ligações.

De olho nisso, Crawford (2014) defende que a autenticação em dispositivos móveis não deve ser pensada da mesma forma que nos desktops e laptops. Para ele, a principal diferença está na forma como os usuários interagem com esses dispositivos. Crawford (2014) entende que os usuários utilizam os dispositivos móveis em "rajadas", buscando-os com maior frequência, mas por curtos períodos de tempo. Nesse caso, quando os usuários são obrigados a utilizar mecanismos de autenticação impróprios para dispositivos móveis, eles podem ser importunados ao ponto de desativar a autenticação ou usar credenciais inseguras. Just (2014) reforça essa visão e propõe que a duração e a frequência da autenticação dos usuários em dispositivos móveis sejam consideradas no projeto de mecanismos de autenticação. Ele acredita que esses fatores podem ser ajustados para que os usuários gastem menos tempo autenticando e isso pode tornar os usuários menos resistentes a utilização de mecanismos de autenticação em dispositivos móveis.

Uma das linhas de pesquisa que se destaca nesse contexto é a autenticação implícita (JAKOBSSON *et al.*, 2009) ou autenticação transparente e contínua (CRAWFORD, 2014). Nessa linha, os mecanismos de autenticação buscam exigir menor participação ativa do usuário e realizar autenticação continuamente a partir de traços biométricos, fisiológicos e comportamentais, que podem ser obtidos de forma transparente pelos dispositivos móveis a partir de dados que o usuário gera quando interage ou simplesmente transporta esses dispositivos. O estudo recente realizado por Patel *et al.* (2016) apresenta uma visão geral de diversas propostas nessa linha de pesquisa. Alguns exemplos dos dados utilizadas nesses trabalhos para a obtenção de traços biométricos são localizações, toques na tela, teclas pressionadas, textos digitados, marcha, face, voz, ruídos sonoros, aplicações usadas, sítios web acessados, ligações e mensagens de texto recebidas e enviadas.

Para oferecer mais segurança, as propostas de mecanismos de autenticação transparente e contínua utilizam sistemas multibiométricos (JAIN *et al.*, 2007) (*i.e.* multimodal),

combinando dois ou mais traços biométricos para construir o perfil biométrico do usuário do dispositivo móvel. Um perfil biométrico é formado de padrões reconhecidos a partir de *features* extraídas de amostras de traços biométricas coletadas de um usuário autorizado. Assim, o desempenho do sistema biométrico pode mudar de acordo com as *features* utilizadas.

Para este trabalho, o conceito de desempenho envolve qualquer aspecto de um sistema computacional que possa ser avaliado por meio de medidas[1] mensuráveis, assim como para Jain (1990). Para ele, as medidas representam os critérios usados para comparar o desempenho durante uma avaliação. Além disso, para Jain (1990), uma avaliação pode ser considerada um *benchmarking* se forem utilizados conjuntos de dados representativos para a realização dos experimentos. Nesse caso, esses dados são chamados de *benchmarks*.

O foco deste trabalho está nas medidas relacionados à eficácia e eficiência do reconhecimento de padrões a partir de *features* biométricas.

Nesse contexto, a eficácia diz respeito à produção do resultado esperado com ausência de erros (MICHAELIS, 2015). Essa definição está relacionada com as medidas de confiabilidade na avaliação de desempenho (JAIN, 1990), que buscam classificar os erros possíveis em um sistema computacional e determinar as suas respectivas probabilidades de ocorrência. No aprendizado de máquina, as medidas de desempenho mais utilizadas para avaliar algoritmos de classificação envolvem o erro (ALPAYDIN, 2014).

Já a eficiência diz respeito à utilização racional dos recursos computacionais para a produção dos resultados esperados, seja na avaliação de sistemas computacionais (JAIN, 1990), como também no projeto de algoritmos (CORMEN, 2009) e no aprendizado de máquina (ALPAYDIN, 2014). Assim, é considerado mais eficiente o sistema ou algoritmo que utiliza menos recursos computacionais para produzir o mesmo resultado. Alguns exemplos de recursos computacionais considerados em avaliações de desempenho são o uso de CPU, memória principal, memória de armazenamento e bateria.

1.2 Motivação

Segundo Jain *et al.* (2007), um sistema biométrico é essencialmente um sistema de reconhecimento de padrões. De acordo com Alpaydin (2014), na engenharia, as pesquisas

[1] Jain (1990) utiliza mais frequentemente o termo métrica (*metric*) no contexto da avaliação de desempenho. Já Alpaydin (2014) prefere usar o termo medida (*measure*) no contexto de aprendizado de máquina. No entanto, ambos os autores usam os dois termos citados como sinônimos. Nesta dissertação, que envolve experimentos de aprendizado de máquina, utiliza-se o termo medida ao longo do texto.

no campo de reconhecimento de padrões estão relacionadas com a inteligência artificial, mais especificamente com os algoritmos de aprendizado de máquina supervisionado e classificação. De fato, Patel *et al.* (2016) e Khan (2016) identificaram em seus trabalhos um grande número de propostas de mecanismos de autenticação transparente e contínua existentes na literatura que estão apoiadas em técnicas de reconhecimento de padrões a partir de *features* biométricas e utilizam amplamente algoritmos de aprendizado de máquina.

Segundo Domingos (2012), no aprendizado de máquina, as *features* são construídas em um processo chamado de engenharia de *features*. Esse processo envolve tentativa e erro e é comumente a atividade que demanda a maior parte dos esforços de um projeto de aprendizado de máquina. Ele considera que a aprendizagem pode ser fácil se você utilizar as *features* certas. Além disso, ele considera também que o conjunto de *features* utilizado representa o fator mais importante para o sucesso de aplicações de aprendizado de máquina.

Por isso, assim como no campo do aprendizado de máquina, pode-se considerar que a escolha do conjunto de *features* biométricas é a chave para o desempenho dos mecanismos de autenticação transparente e contínua. No entanto, segundo Marcel (2013), o desempenho de sistemas biométricos é muitas vezes desconhecido ou impossível de ser comparado, e a falha está na avaliação. Para ele, a medição do desempenho de sistemas biométricos não é realizada adequadamente por falta de padronização dos métodos de avaliação utilizados. Nesse sentido, Marcel (2013) considera que as iniciativas existentes estão focando apenas em algumas poucas modalidades biométricas específicas. Além disso, Khan (2016) constatou também que vários esquemas proeminentes de autenticação transparente e contínua apresentam um desempenho ruim quando são avaliados em condições operacionais realistas.

Assim, Marcel (2013) considera que é necessário estabelecer um *framework* para sistematizar a avaliação das tecnologias biométricas, definindo para isso medidas e critérios adequados.

Nesta pesquisa foram encontrados na literatura indícios de deficiências na reprodutibilidade (VANDEWALLE *et al.*, 2009), granularidade (JAIN *et al.*, 2007) e análise estatística dos resultados dos experimentos (JAIN, 1990) realizados para a avaliação de propostas de mecanismos de autenticação transparente e contínua.

Sobre a reprodutibilidade, as deficiências dizem respeito ao detalhamento insuficiente das descrições das *features* utilizadas, a condução de avaliação não padronizada e a indisponibilidade dos dados utilizados.

Sobre a granularidade, as deficiências estão ligadas à avaliação de sistemas multibiométricos como um todo, sem considerar o desempenho individual gerado pelas *features* biométricas propostas para cada traço biométrico utilizado.

Sobre a análise estatística dos resultados, as deficiências dizem respeito à ausência de evidências do uso de técnicas estatísticas que permitam obter resultados conclusivos sobre o desempenho dos mecanismos propostos.

Com esses problemas, torna-se difícil ou até mesmo impossível a reprodução dos experimentos e a comparação do desempenho entre *features* biométricas equivalentes. Consequentemente, isso impede o reuso das *features* biométricas em novas pesquisas para construção de sistemas com melhor desempenho.

1.3 Objetivo e Metas

O objetivo desta pesquisa é, portanto, propor um processo de avaliação de *features* biométricas para autenticação transparente e contínua em dispositivos móveis, baseado em aprendizado de máquina. Além disso, o processo é aplicado em um estudo de caso para a avaliação de *features* de localização.

Para alcançar o objetivo desta pesquisa, foram estabelecidas as seguintes metas:

- Identificar na literatura os trabalhos relacionados com foco na padronização de métodos para a avaliação biométrica;
- Identificar na literatura boas práticas para a avaliação de desempenho de sistemas computacionais e a realização de experimentos de aprendizado de máquina;
- Elaborar um processo de avaliação de *features* biométricas para autenticação transparente e contínua em dispositivos móveis baseado em aprendizado de máquina; e
- Avaliar o processo proposto com um estudo de caso.
  - Identificar os trabalhos relacionados que utilizam *features* de localização *outdoor* para autenticação transparente e contínua em dispositivos móveis;
  - Analisar os trabalhos relacionados e definir conjuntos de *features* de localização *outdoor* que representam as estratégias usadas na literatura; e
  - Avaliar o desempenho dos conjuntos de *features* de localização *outdoor* definidos usando o processo proposto.

Para a avaliação do processo proposto foi mantido o foco nas *features* biométricas

extraídas de dados de localização *outdoor* de usuários de dispositivos móveis. No entanto, o processo proposto neste trabalho é elaborado para ser compatível com a avaliação de quaisquer *features* biométricas usadas para autenticação transparente e contínua em dispositivos móveis.

1.4 Resultados Esperados

Assim, os resultados esperados deste trabalho são:

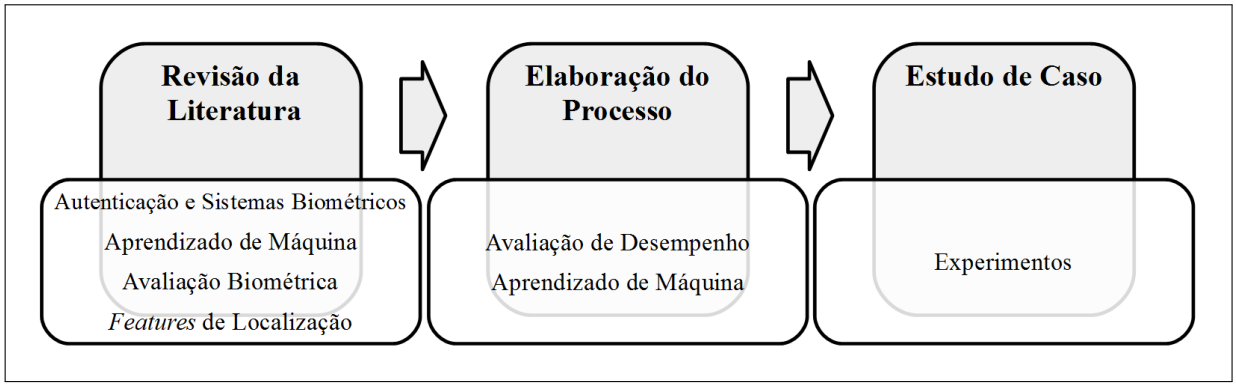
- Um processo baseado em aprendizado de máquina para a avaliação de *features* biométricas utilizadas nos mecanismos de autenticação transparente e contínua em dispositivos móveis; e
- Resultados reutilizáveis obtidos a partir da realização de um estudo de caso envolvendo a eficácia e a eficiência das *features* de localização *outdoor* definidas nesta pesquisa com base nos trabalhos relacionados existentes na literatura.

Espera-se com os resultados desta dissertação apoiar novas pesquisas na área de autenticação transparente e contínua.

1.5 Metodologia

A Figura 1 apresenta a metodologia científica utilizada para alcançar os objetivos específicos deste trabalho. Os detalhes da metodologia utilizada em cada uma dessas etapas estão descritos no restante desta seção.

Figura 1 – Metodologia de trabalho utilizada nesta pesquisa.



Fonte – O autor.

Neste trabalho é realizada uma revisão da literatura sobre os principais conceitos e definições nas áreas de aprendizado de máquina, autenticação e biometria. O resultado dessa

parte da revisão é apresentado no Capítulo 2 deste trabalho.

É realizada também uma revisão da literatura para identificar os trabalhos relacionados sobre avaliação biométrica e seus processos, que é o foco principal deste trabalho. Além disso, são identificados também trabalhos baseados em *features* de localização *outdoor* para a avaliação do processo proposto. Os detalhes dos trabalhos relacionados são descritos no Capítulo 3 deste trabalho.

Para identificação dos trabalhos relacionados baseados em *features* de localização *outdoor* é seguida uma parte do protocolo sistemático de busca proposto por Kitchenham e Charters (2007). Na fase de planejamento dessa revisão são definidos string de busca, critérios de seleção e as fontes de pesquisa. Em seguida, na fase de execução são conduzidas as buscas e realizada a filtragem dos resultados. Por fim, na fase de análise dos resultados, os trabalhos são detalhadamente analisados, especialmente no que diz respeito às *features* de localização *outdoor*, estratégia de aprendizado e medidas de desempenho utilizadas.

Para a elaboração do processo de avaliação de *features* biométricas para autenticação transparente e contínua em dispositivos móveis são utilizadas as etapas da abordagem sistemática para a avaliação de desempenho de sistemas computacionais proposta por Jain (1990) e também as diretrizes propostas por Alpaydin (2014) para projeto e análise de experimentos de aprendizado de máquina.

O objetivo deste trabalho envolve a avaliação de desempenho de sistemas computacionais e o aprendizado de máquina a partir de *features* biométricas. Assim, as obras de Alpaydin (2014) e Jain (1990) foram escolhidas como referências principais para este trabalho porque abordam com profundidade os temas necessários e são reconhecidas nas suas respectivas áreas. As etapas da abordagem sistemática propostas por Jain (1990) buscam evitar os erros mais recorrentes dos projetos de avaliação de desempenho de sistemas computacionais. Já as diretrizes propostas por Alpaydin (2014) buscam orientar a realização de experimentos de aprendizado de máquina de acordo com o método estatístico que apoia a pesquisa científica em várias áreas. Assim, essas diretrizes permitem projetar experimentos corretamente e analisar os dados coletados de forma a serem capazes de compreender a influência de fatores controláveis, identificar a expectativa de erro experimental e, por fim, extrair resultados conclusivos com nível de confiança definido e aceitável. Os detalhes do processo proposto são descritos no Capítulo 4 deste trabalho.

Como forma de avaliar o processo proposto neste trabalho, um estudo de caso é

conduzido com experimentos para avaliação da eficácia e eficiência das *features* de localização *outdoor* definidas a partir dos trabalhos relacionados identificados durante a revisão da literatura. Os detalhes dos experimentos realizados são descritos no Capítulo 5 deste trabalho.

1.6 Organização da Dissertação

Esta dissertação está organizada em cinco capítulos. O presente capítulo apresenta uma introdução ao tema, descrevendo a contextualização, a motivação, o problema, os objetivos, os resultados esperados e a metodologia utilizada.

No Capítulo 2 são apresentados os principais conceitos relacionados à autenticação de usuários e aprendizado de máquina, focando nos detalhes da avaliação de sistemas biométricos e do design de experimentos de aprendizado de máquina.

No Capítulo 3 são apresentados os trabalhos relacionados, destacando as principais semelhanças e diferenças com esta pesquisa. Além disso, são descritos os detalhes metodológicos relacionados à revisão da literatura realizada.

No Capítulo 4 são descritos os detalhes do processo proposto para avaliação de *features* biométricas.

No Capítulo 5 são apresentados os detalhes do estudo de caso da avaliação do processo proposto, realizada com foco nas *features* de localização *outdoor*, incluindo os resultados obtidos nos experimentos.

Por fim, o Capítulo 6 apresenta as conclusões desta pesquisa e os possíveis trabalhos futuros identificados.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são abordados os principais conceitos e definições que serviram de base para o desenvolvimento desta pesquisa. As próximas seções deste capítulo estão estruturadas da seguinte forma: a seção 2.1 aborda a autenticação de usuários; na sequência, a seção 2.2 aborda o aprendizado de máquina; e, por fim, a seção 2.3 apresenta as considerações finais.

2.1 Autenticação de Usuários

Nesta seção serão abordados os conceitos gerais relacionados a autenticação de usuários com foco em biometria e autenticação transparente e contínua.

2.1.1 Visão Geral

Para Clarke (2011), os principais objetivos da segurança da informação são confidencialidade, integridade e disponibilidade. Para ele, esses objetivos são apoiados por serviços essenciais como autenticação, autorização e responsabilização (*Authentication, Authorization, and Accountability* – AAA). Assim, para garantir confidencialidade e integridade das informações, é necessário que os sistemas computacionais confirmem a identidade dos seus usuários. Só assim as permissões apropriadas de acesso podem ser concedidas, evitando que usuários não autorizados leiam ou modifiquem indevidamente informações. Por isso, Clarke (2011) considera a autenticação fundamental para a manutenção da segurança dos sistemas computacionais.

Clarke (2011), assim como Nakamura e Geus (2007), dividem as diversas abordagens de autenticação existentes em três categorias, observando exclusivamente a natureza do elemento secreto utilizado. São elas:

- Algo que você sabe – abordagens baseadas em elemento de informação de conhecimento exclusivo do usuário legítimo do sistema. Exemplos: senhas numéricas (*Personal Identification Number* – PIN) e senhas alfanuméricas (*password*);
- Algo que você tem – abordagens baseadas em elemento físico de posse exclusiva do usuário legítimo do sistema. Exemplos: *tokens* de segurança e cartões inteligentes (*smartcards*); e
- Algo que você é – abordagens baseada em *features* biométricas exclusivas do usuário legítimo do sistema. Exemplos: sistemas de reconhecimento de impressão digital (*fingerprint*) e facial.

Essas categorias de autenticação representam de forma geral os mecanismos adotados pelos sistemas. Entretanto, durante a fase exploratória desta pesquisa foi possível perceber que a autenticação biométrica tem se mostrado especialmente importante para as novas propostas envolvendo a autenticação de usuários em dispositivos móveis (PATEL *et al.*, 2016)(KHAN, 2016). Por isso, neste trabalho, será dado foco especial na biometria.

2.1.2 Biometria

Para Clarke (2011), a definição moderna de biometria vai mais longe do que simplesmente referir-se a uma característica única do usuário. Segundo ele, o *International Biometrics Group* (IBG) define biometria como o uso automatizado de *features* fisiológicas ou comportamentais para determinar ou verificar a identidade do usuário.

Nessa definição o primeiro ponto a destacar é o termo "*features*". Como já foi antecipado na seção 1.1, as *features* biométricas são extraídas de amostras de traços biométricos humanos. Todos os usuários que precisam ser reconhecidos por meio da biometria devem possuir esse traço biométrico, mas o que diferencia esses usuários são as *features* biométricas extraídas.

Para entender melhor o conceito de *feature* biométrica, pode-se tomar como exemplo um cenário envolvendo a face como traço biométrico. Nesse caso, a amostra gerada do traço biométrico é representada por uma imagem do rosto do usuário. Já as *features* devem ser extraídas dessa imagem e podem ser definidas como as distâncias entre os olhos, nariz, boca e orelhas. Porém, as *features* podem ser definidas de outras formas, inclusive como o próprio dado bruto (*i.e. raw data*) da imagem, representado nesse caso, por exemplo, por um matriz de bits (*i.e. bitmap*). Assim, um sistema biométrico utiliza as *features* para realizar a autenticação dos usuários.

Outros pontos a destacar nessa definição de biometria apresentada no início desta seção são os termos "fisiológicas" e "comportamentais". Segundo Jain *et al.* (2007), as *features* biométricas podem ser extraídas a partir de traços fisiológicos ou comportamentais. Os traços fisiológicos estão relacionados a aspectos físicos do corpo humano. Alguns exemplos de traços fisiológicos são a impressão digital, a face, a íris, as mãos e as orelhas. Já os traços comportamentais estão relacionados a padrões de comportamento apresentados pelo usuário. Alguns exemplos de traços comportamentais são a voz, a assinatura, a marcha, a digitação e a localização.

Mais um ponto a destacar nessa definição de biometria é o termo "automatizado".

Embora existam inúmeros traços humanos de onde podem ser extraídos conjuntos singulares de *features* dos indivíduos, essas *features* só podem ser consideradas biométricas, segundo essa definição, uma vez que o processo de autenticação a partir delas possa ser realizado de forma automatizada. Por exemplo, Clarke (2011) cita que o DNA é provavelmente o traço de onde pode ser extraído o conjunto singular de *features* mais conhecido pela ciência. No entanto, atualmente, essas *features* não podem ser qualificadas como biométricas, porque não existe ainda um processo automatizado para identificar os indivíduos baseado nelas.

Por fim, o último ponto a destacar nessa definição é o trecho "determinar ou verificar a identidade". Segundo Jain *et al.* (2007), os sistemas biométricos podem operar de dois modos: verificação e identificação.

A verificação, também chamada de autenticação, é o processo de confirmação que o utilizador é o usuário autorizado do sistema que ele alega ser. Um sistema biométrico em modo de verificação tem funcionamento muito similar às técnicas de autenticação baseadas em senhas utilizadas em sistemas computacionais. Nelas o usuário digita um nome de usuário e uma senha, reivindicando assim uma identidade. Os sistemas biométricos funcionam de forma semelhante quando operam em modo de verificação, só que a senha é substituída por *features* biométricas. Para verificar a autenticidade do usuário, o sistema apenas compara as *features* biométrica fornecidas com as *features* biométricas registradas no banco de dados biométricos do sistema e associadas à identidade reivindicada.

Já no modo de identificação o usuário não alega ser alguém, pois simplesmente fornece suas *features* biométricas ao sistema. Cabe ao sistema biométrico, nesse caso, determinar se as *features* fornecidas são compatíveis com algum dos usuários autorizados do sistema. Para isso, é preciso comparar as *features* biométricas apresentadas com todos os usuários registrados no banco de dados biométrico.

Assim, esses dois modos de operação dos sistemas biométricos representam problemas muito diferentes. A identificação é necessária apenas em cenários onde se têm múltiplos usuários. Além disso, os recursos computacionais são mais exigidos no modo de identificação. Baseado nisso e também levando em consideração a dinâmica de uso dos dispositivos móveis e a limitação computacional comumente encontrada neles, o foco deste trabalho será nos sistemas biométricos operando em modo de verificação.

2.1.2.1 Requisitos Biométricos

É importante lembrar que cada traço biométrico tem pontos fortes e fracos. Assim, a escolha de um traço biométrico para uma aplicação pode depender de várias requisitos. Segundo Jain *et al.* (2007), os principais requisitos que definem a adequação de um traço fisiológico ou comportamental para uma aplicação biométrica são:

- Universalidade. Os usuários do sistema devem possuir o traço biométrico;
- Unicidade. O traço biométrico deve variar suficientemente entre os usuários do sistema;
- Permanência. O traço biométrico de cada usuário do sistema deve ser suficientemente invariante ao longo do tempo;
- Mensurabilidade. O traço biométrico deve permitir a aquisição suficientemente fácil de dados brutos e a extração de *features* representativas utilizando dispositivos adequados sem causar inconveniente indevido aos usuários do sistema;
- Desempenho. A acurácia do reconhecimento de padrões e os recursos computacionais utilizados devem atender às restrições do sistema;
- Aceitabilidade. Os usuários do sistema devem estar dispostos a apresentar o traço biométrico para a aquisição pelo sistema; e
- Circunvenção. o traço biométrico deve ser suficientemente difícil de imitar usando artefatos físicos, no caso de traços fisiológicos, ou usando mimetismo, no caso de traços comportamentais.

Jain *et al.* (2007) esclarecem que não é esperado que um traço biométrico satisfaça completamente todos esses requisitos, pois nenhum traço biométrico é ideal. O nível desejável para cada um desses requisitos é estabelecido dependendo da natureza e dos requisitos da aplicação, e das propriedades das *features* biométricas.

2.1.2.2 Sistemas Biométricos

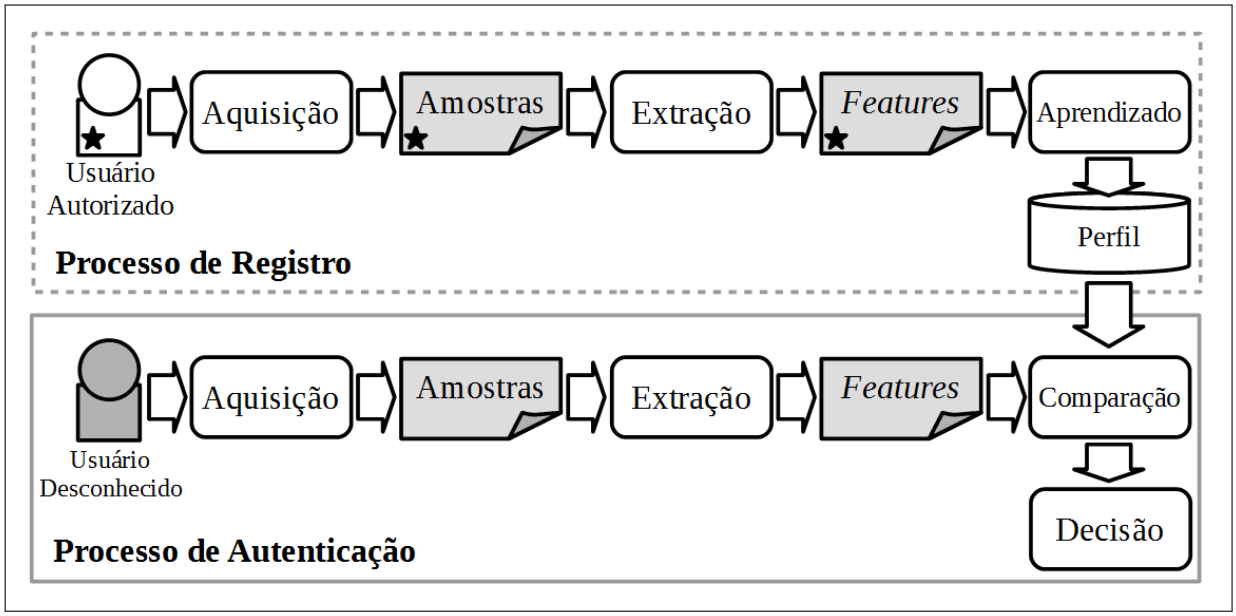
Com base em Clarke (2011) e Jain *et al.* (2007), um sistema biométrico operando em modo de verificação pode ser representado principalmente pelos seguintes processos:

- Aquisição de Amostra. O processo de aquisição obtém dados de traços biométricos do usuário. Esse processo busca garantir a qualidade dos dados sem causar excesso de inconveniência para o usuário durante a coleta. O resultado do

processo de aquisição é um conjunto de amostras;

- Extração de *Features*. O processo de extração realiza a redução da grande quantidade de dados contidos nas amostras em uma pequena quantidade de dados que são realmente úteis, garantindo perda mínima de informação. O resultado do processo de extração é um conjunto de *features* biométricas;
- Aprendizado de Padrões. O processo de aprendizado realiza o reconhecimento de padrões a partir das *features* biométricas. O resultado do processo de aprendizado é um perfil biométrico que será armazenado no banco de dados biométrico;
- Comparação. O processo de comparação confronta os padrões reconhecidos e armazenados no perfil biométrico da identidade reivindicada com as *features* biométricas extraídas das amostras do usuário atual do sistema. O resultado do processo de comparação é aceitar ou rejeitar o usuário como autêntico, com base na compatibilidade entre o perfil armazenado e as *features*, considerando um determinado limite de tolerância; e
- Registro. O processo de registro realiza o cadastramento prévio dos usuários autorizados no sistema. A linha pontilhada na Figura 2 significa que o processo de registro é realizado com menor frequência do que o processo de autenticação.

Figura 2 – Modelo simplificado de sistema biométrico.



Fonte – O autor, baseado em Clarke (2011) e Jain *et al.* (2007).

2.1.2.3 Multibiometria

Segundo Jain *et al.* (2007), os sistemas biométricos usam comumente apenas um traço biométrico para estabelecer a identidade e são considerados sistemas uni-biométricos. No entanto, ele considera que algumas das limitações desses sistemas podem ser resolvidas ou atenuadas com sistemas conhecidos como multibiométricos. Esses sistemas são projetados para combinar múltiplas fontes da informação biométrica. De acordo com a natureza dessas fontes, um sistema multibiométrico pode ser classificado em uma das seguintes categorias (JAIN *et al.*, 2007):

- Multi-sensor. Empregam vários sensores para capturar um único traço biométrico do usuário;
- Multi-algoritmo. Consolidam a saída de múltiplos algoritmos de extração de *features* ou múltiplas comparações operando no mesmo conjunto de *features*;
- Multi-instância ou Multi-unitários. Usam múltiplas instâncias do mesmo traço biométrico (*e.g.* polegar direito e esquerdo; íris direita e esquerda);
- Multi-amostra. Adquirem múltiplas amostras do mesmo traço biométrico.
- Multi-modo. Estabelecem identidade com base em *features* extraídas a partir de múltiplos traços biométricos.
- Híbrido. Descrevem sistemas que integram um subconjunto dos cinco cenários apresentados anteriormente.

2.1.2.4 Avaliação de Sistemas Biométricos

Para Phillips *et al.* (2000), a avaliação de sistema biométrico serve para mensurar a adequação desses sistemas aos requisitos da aplicação. Segundo eles, normalmente, as avaliações de sistemas biométricos exigem que um avaliador independente faça o planejamento, coleta de dados, execução dos experimentos e análise dos resultados. Para que uma avaliação seja aceita pela comunidade biométrica, Phillips *et al.* (2000) recomendam que os detalhes do procedimento de avaliação sejam publicados juntamente com o protocolo de avaliação, procedimentos de teste, resultados de desempenho e exemplos representativos do conjunto de dados. Além disso, as informações sobre a avaliação e os dados devem ser suficientemente detalhados para que usuários, desenvolvedores e fornecedores possam reproduzir a avaliação. Vandewalle *et al.* (2009) também pesquisaram sobre a reprodutibilidade das pesquisas na área de processamento de sinais

e acreditam que resultados reproduzíveis aumentam o impacto da pesquisa e, consequentemente, o número de citações e a utilização desses resultados em outras pesquisas, aplicações comerciais e aulas.

No entanto, analisando os trabalhos publicados na área de autenticação transparente e contínua, é possível perceber que os pesquisadores avaliam as suas próprias técnicas propostas e não seguem essas recomendações de Phillips *et al.* (2000). Marcel (2013) também percebe deficiências na avaliação dos sistemas biométricos e considera que o desempenho desses sistemas é muitas vezes desconhecido ou impossível de comparar. Além disso, Khan (2016) constatou que em condições operacionais realistas, vários esquemas de autenticação implícita proeminentes não fornecem segurança adequada, pois apresentam um desempenho ruim quando são avaliados. Quando a avaliação é realizada pelos próprios autores de uma proposta a reprodutibilidade dos resultados obtidos por essas avaliações é ainda mais importante. Nesses casos os resultados precisam ser reproduzidos, verificados por outros pesquisadores e comparados com outros trabalhos para que as técnicas propostas sejam reutilizadas ou melhoradas. Além disso, o processo de avaliação comumente utilizado nos trabalhos relacionados deve ser seguido ao máximo, incluindo as medidas selecionadas. Isso faz com que os resultados obtidos possam ser comparados.

Na avaliação de sistemas biométricos, o desempenho de um sistema operando em modo de verificação é tradicionalmente caracterizado por duas medidas de erro: a taxa de rejeição falsa (*i.e. False Rejection Rate* - FRR); e a taxa de aceitação falsa (*i.e. False Acceptance Rate* - FAR) (CLARKE, 2011). A rejeição falsa ocorre quando um sistema rejeita uma identidade válida e a aceitação falsa ocorre quando um sistema aceita uma identidade não válida (CLARKE, 2011). Em um sistema biométrico ideal ambas as taxas de erro são iguais a zero. Porém, os sistemas biométricos não são perfeitos (PHILLIPS *et al.*, 2000). Além disso, as taxas de erro dos sistemas biométricos são relacionadas e ajustáveis (PHILLIPS *et al.*, 2000). Isso quer dizer que quando um sistema é ajustado para negar acesso a todos os usuários, a FRR é 100% e a FAR é 0%. No outro extremo, quando o sistema é ajustado para conceder acesso a todos, a FRR é 0% e a FAR é 100%. Na prática, os sistemas biométricos operam entre esses dois extremos e para a maioria das aplicações é necessário ajustar um parâmetro do sistema para obter a FAR desejada (PHILLIPS *et al.*, 2000). Esse ajuste resulta automaticamente em uma FRR correspondente. Assim, aumentar o nível de segurança do sistema (FAR menor) significa também diminuir o nível de usabilidade (FRR maior), e vice-versa (NAKAMURA; GEUS, 2007).

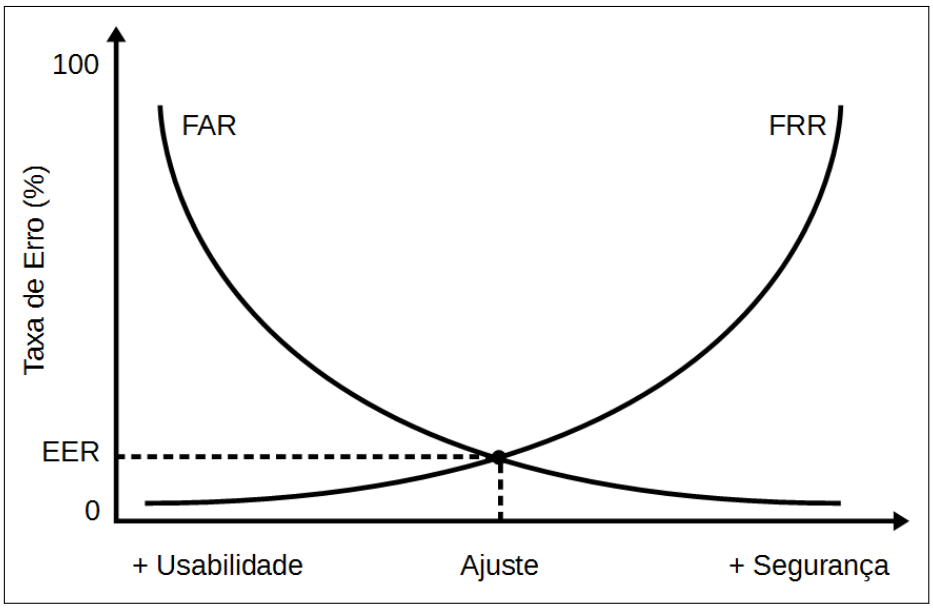
As taxas FAR e FRR representam medidas gerais e incluem os erros contabilizados em um sistema biométrico como um todo (CLARKE, 2011). Além delas, existem também outras quatro medidas usadas para contabilizar erros específicos dos processos de aquisição, extração, registro e comparação de sistemas biométricos: a falha na aquisição (*i.e. Failure To Acquire* - FTA); a falha no registro (*i.e. Failure To Enroll* - FTE); a taxa de correspondência falsa (*i.e. False Match Rate* - FMR); e a taxa de não correspondência falsa (*i.e. False Non-Match Rate* - FNMR) (CLARKE, 2011). A FTA mensura a taxa de erros específica nos processos de aquisição de amostras e extração de *features* (CLARKE, 2011). A FTE contabiliza a taxa de erros específica no processo de registro de usuários (CLARKE, 2011). Por fim, as FMR e FNMR medem respectivamente as taxas de aceitação falsa e rejeição falsa específicas do processo de comparação (CLARKE, 2011). A principal diferença entre as medidas FMR/FNMR e FAR/FRR é que estas últimas incluem erros que ocorrem em todos os processos de um sistema biométrico.

As taxas de erro descritas nesta subseção são acompanhadas por duas estatísticas de acerto: a taxa de rejeição verdadeira (*i.e. True Reject Rate* - TRR); e a taxa de aceitação verdadeira (*i.e. True Accept Rate* - TAR) (CLARKE, 2011). A rejeição verdadeira ocorre quando um sistema rejeita corretamente uma identidade não válida e a aceitação verdadeira ocorre quando um sistema aceita corretamente uma identidade válida (CLARKE, 2011).

Segundo Jain *et al.* (2007), os fornecedores biométricos comumente informam o desempenho dos seus sistemas em termos de FAR e FRR. No entanto, para Phillips *et al.* (2000), como os parâmetros dos sistemas podem ser ajustados para obter diferentes valores para FAR, torna-se frequentemente difícil comparar sistemas que apresentam medições de desempenho com base em diferentes valores de FAR. Por isso, outra forma comumente usada para a comparação de sistemas biométricos é a taxa de erro igual (*i.e. Equal Error Rate* - EER) (CLARKE, 2011). Conforme apresentado na Figura 3, essa medida corresponde ao ponto onde as curvas FAR e FRR se cruzam e apresentam valores iguais (CLARKE, 2011). Essas curvas são plotadas por meio de medições das respectivas taxas em vários cenários de ajuste operacional do sistema (JAIN *et al.*, 2007). Uma EER menor significa um sistema biométrico melhor e com menor taxa de erro. É possível constatar que o EER é muito utilizado na avaliação de várias das propostas de mecanismos de autenticação transparente e contínua analisadas nos trabalhos de Patel *et al.* (2016) e Khan (2016).

Outro tipo de curva utilizada na avaliação e comparação de sistemas biométricos é a curva ROC (*i.e. Receiver Operating Characteristic* - ROC) (CLARKE, 2011). Essa curva pode

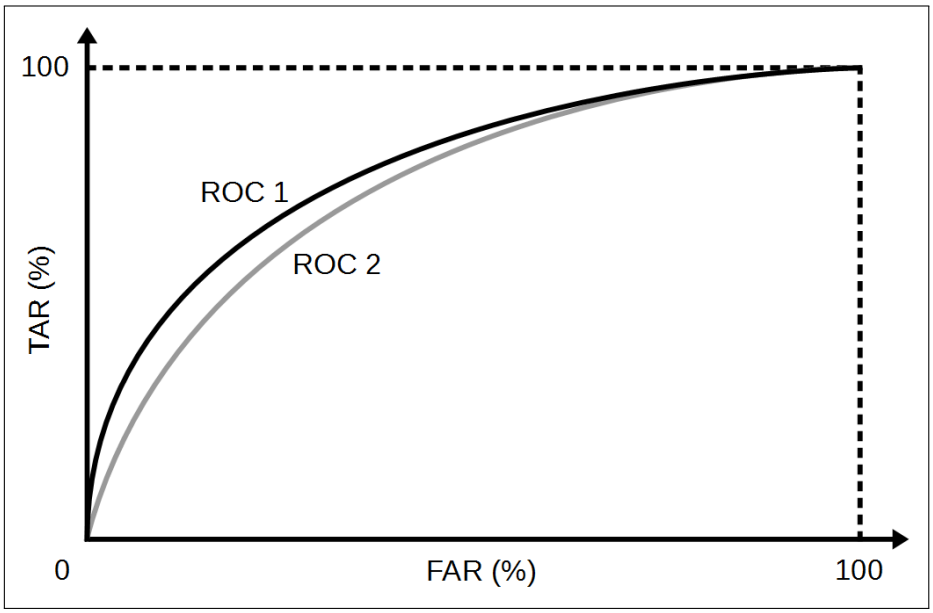
Figura 3 – Curvas FAR e FRR e a medida EER.



Fonte – O autor, traduzido de Clarke (2011).

ser plotada por meio de medições das taxas FAR e TAR obtidas em vários cenários de ajuste operacional do sistema (JAIN *et al.*, 2007). A curva ROC serve para auxiliar a comparação gráfica das avaliações de sistemas biométricos (JAIN *et al.*, 2007). Na Figura 4, a curva ROC 1 tem uma maior área sob a curva, significando um TAR médio maior. Por isso, essa curva representa um sistema com melhor desempenho.

Figura 4 – Curvas ROC para comparação de sistemas biométricos.



Fonte – O autor, traduzido de Clarke (2011).

Segundo Clarke (2011), essas taxas de erro e acerto são médias dos resultados obtidos nos testes com amostras da população. Portanto, elas representam o resultado esperado que um usuário pode alcançar (CLARKE, 2011). No entanto, os resultados individuais podem variar dependendo de características particulares das amostras selecionadas para a avaliação (CLARKE, 2011).

No geral, a biometria comportamental apresenta taxa de acerto inferior quando comparada com a biometria fisiológica (CLARKE, 2011). Por outro lado, a biometria comportamental tende a ser mais transparente e conveniente para o usuário (CLARKE, 2011) e isso facilita a sua utilização na autenticação transparente e contínua.

2.1.3 Autenticação Transparente e Contínua

A autenticação do usuário é tradicionalmente realizada pelos sistemas utilizando uma abordagem de ponto de entrada (CLARKE, 2011). Essa abordagem é comumente caracterizada por uma interface que força a autenticação do usuário antes que ele tenha acesso ao sistema. Quando autenticado com sucesso, o usuário tem acesso ao sistema, ou, pelo menos, a um subconjunto de suas funções, por um período de tempo sem a necessidade de realizar nova autenticação. Em algumas aplicações, o sistema é bloqueado depois de um tempo pré-determinado de inatividade do usuário. Já em outras, o sistema simplesmente permanece pronto para o uso até que o usuário indique explicitamente a intenção de bloqueá-lo ou sair.

Com um número cada vez maior de "coisas"ao redor dos usuários e, ao mesmo tempo, interações mais rápidas e frequentes com esses dispositivos, surgem novas abordagens para autenticação dos usuários. Crawford (2014) acredita que os dispositivos móveis dão origem a uma necessidade de autenticação sem a participação ativa do usuário, ou, pelo menos, com um envolvimento muito limitado dele. Na tentativa de superar as limitações impostas pela utilização de técnicas tradicionais de autenticação nos dispositivos móveis, alguns autores têm proposto o uso de autenticação implícita (JAKOBSSON *et al.*, 2009), transparente e contínua (CLARKE, 2011).

Para Jakobsson *et al.* (2009), autenticação implícita consiste na autenticação dos usuários com base nas ações que eles já realizam enquanto interagem com os dispositivos móveis. Segundo os autores, a autenticação implícita pode ser implementada por qualquer tipo de computador, mas é particularmente adequada para dispositivos móveis, pois eles comumente possuem múltiplos sensores que são úteis para identificação precisa dos usuários. Além disso,

Chow *et al.* (2010) identificaram que a combinação dos dados provenientes desses diversos sensores permite elevar o grau de certeza sobre a autenticidade dos usuários.

Para Clarke (2011), a autenticação transparente e contínua é a autenticação dos usuários baseada em fatores biométricos que minimizam a participação ativa do usuário e proporciona segurança durante todo o período de utilização. Assim, é possível perceber que os conceitos de autenticação implícita, transparente e contínua estão relacionados, pois é natural que um mecanismos de autenticação implícito seja também transparente. Ao mesmo tempo, um mecanismo de autenticação contínua precisa ser também implícito e transparente. Caso não seja, o mecanismo demandará mais interações explícitas do usuário para autenticação, tornando-o impraticável.

Crawford (2014) defende que a autenticação em dispositivos móveis não deva ser pensada da mesma forma que nos *desktops* e *laptops*. Para ele, a principal diferença está na forma como os usuários interagem com esses dispositivos. Crawford entende que os usuários utilizam os dispositivos móveis em "rajadas", buscando-os com maior frequência, mas por curtos períodos de tempo. Nesse caso, quando os usuários são obrigados a utilizar mecanismos de autenticação impróprios para dispositivos móveis, eles podem ser importunados ao ponto de desativar a autenticação ou usar credenciais inseguras. Just (2014) reforça essa visão e propõe que a duração e a frequência da autenticação dos usuários em dispositivos móveis sejam consideradas no projeto de mecanismos de autenticação. Ele acredita que esses fatores possam ser ajustados para que os usuários gastem menos tempo se autenticando e isso pode diminuir a resistência desses usuários contra mecanismos de autenticação.

Segundo Khan (2016), a autenticação transparente e contínua (ou autenticação implícita como o referido autor prefere chamar juntando os dois termos) requer aprendizado de padrões distintos de uso do dispositivo. Para ele, o padrão de uso atual do dispositivo precisa ser comparado com o padrão do proprietário, aprendido anteriormente, na tentativa de detectar comportamentos anormais. Nesse contexto, Khan (2016) esclarece que as técnicas de aprendizado de máquina são utilizadas para o aprendizado e a verificação desses padrões.

2.2 Aprendizado de Máquina

Nesta seção serão abordados os conceitos gerais relacionados ao Aprendizado de Máquina com foco na avaliação de desempenho por meio de Experimentos.

2.2.1 Visão Geral

Na computação, os algoritmos são sequências bem definidas de instruções que usam um conjunto de entradas para produzir um conjunto de saídas necessárias para resolver um problema computacional específico (CORMEN, 2009). No entanto, não é possível criar algoritmos para alguns problemas computacionais quando não se sabe como transformar as entradas nas saídas necessárias para a solução (ALPAYDIN, 2014). Em alguns desses casos, ao mesmo tempo que falta conhecimento para a construção do algoritmo, sobram amostras a partir das quais é possível compreender a relação entre as entradas possíveis e as saídas necessárias.

Mitchell (1997) define aprendizado de máquina como qualquer programa de computador que melhore o seu desempenho em alguma tarefa por meio da experiência. Assim, em alguns desses casos de problemas sem solução, o aprendizado de máquina pode ser utilizado para extrair automaticamente um algoritmo a partir de um conjunto adequado de amostras (ALPAYDIN, 2014).

A partir da definição de aprendizado de máquina mencionada anteriormente é possível compreender a importância da avaliação para a verificação da melhoria do desempenho e, consequentemente, do aprendizado. O termo desempenho é utilizado em alguns casos com sentido restrito, fazendo menção à utilização de recursos computacionais. No entanto, segundo Jain (1990), o conceito de desempenho é mais amplo e envolve qualquer aspecto de um sistema computacional que possa ser avaliado por meio de medidas mensuráveis.

No aprendizado de máquina, as *features* são construídas em um processo chamado de engenharia de *features*. Segundo Domingos (2012), esse processo envolve reunir, integrar, limpar e pré-processar dados brutos. Para ele, esse processo envolve tentativa e erro, sendo que o conhecimento específico do domínio dos dados é tão importante quanto o conhecimento técnico sobre os algoritmos de aprendizado de máquina. Ainda segundo Domingos (2012), a engenharia de *features* é a atividade que demanda a maior parte dos esforços de um projeto de aprendizado de máquina. Ele considera isso normal, já que o fator mais importante para o sucesso desses projetos é justamente o conjunto de *features* utilizadas. Para Domingos (2012), a aprendizagem pode ser fácil se as *features* são independentes e diretamente correlacionadas com a classe. Porém, a aprendizagem pode não ser possível se a classe é uma função muito complexa das *features*. Domingos (2012) afirma ainda que, muitas vezes, os dados brutos (*i.e.* raw data) não permitem a aprendizagem, mas é possível extrair *features* a partir deles que oferecem melhor desempenho. Ele acredita que a engenharia de *features* é mais difícil porque é específica do

domínio dos dados, enquanto que as técnicas de aprendizagem já foram bem dominadas e são, em grande número, de propósito geral.

Segundo Alpaydin (2014), o aprendizado de máquina pode ser supervisionado ou não supervisionado. Na aprendizagem supervisionada, o objetivo é aprender o mapeamento entre as entradas e saídas possíveis cujos valores corretos são fornecidos por um supervisor (ALPAYDIN, 2014). Nesse tipo de aprendizagem são usados algoritmos de classificação ou regressão. Os algoritmos de regressão são úteis quando o problema envolve uma função numérica não conhecida e as entradas devem ser mapeadas para um conjunto contínuo de saídas numéricas (ALPAYDIN, 2014). Já os algoritmos de classificação são utilizados quando as entradas devem ser mapeadas para um conjunto discreto de saídas que representam classes (ALPAYDIN, 2014).

Na aprendizagem não supervisionada, não existe supervisor e, por consequência, não existe também um mapeamento correto a priori entre as entradas e saídas (ALPAYDIN, 2014). O objetivo nesse caso é encontrar regularidades na entrada, ou seja, identificar estruturas na entrada de tal forma que certos padrões podem ocorrer com mais frequência do que outros (ALPAYDIN, 2014). Nesse tipo de aprendizagem são utilizados algoritmos de agrupamento (*clustering*). Esses algoritmos são úteis quando é preciso identificar entradas similares e criar um mapeamento para um conjunto de saídas (ALPAYDIN, 2014).

Na autenticação biométrica, o aprendizado de máquina é usado para o reconhecimento de padrões fisiológicos e comportamentais dos usuários (ALPAYDIN, 2014). Um exemplo é o reconhecimento facial. Esta é uma tarefa que os seres humanos sabem fazer sem esforço. Todos os dias as pessoas reconhecem membros da família e amigos, olhando para os seus rostos pessoalmente ou por meio de fotografias, independente de fatores como a pose, iluminação, estilo de cabelo, entre outros. Elas fazem isso inconscientemente e não são capazes de explicar como. Ao mesmo tempo, sabemos que uma imagem de um rosto não é apenas uma coleção aleatória de *pixels*. Um rosto tem estrutura e é simétrico. Assim, o rosto de cada pessoa é um padrão composto por uma combinação particular de olhos, nariz, boca e outros elementos, todos eles localizados em regiões previstas. Analisando amostras de imagens do rosto de uma pessoa, uma máquina pode aprender o seu padrão específico e, em seguida, realizar o reconhecimento quando encontrar esse mesmo padrão em uma imagem (ALPAYDIN, 2014). O desempenho de aprendizagem e reconhecimento desses padrões pode ser avaliado por meio da realização de experimentos.

2.2.2 Experimentos

Em aprendizado de máquina, realizar experimentos significa treinar um modelo de aprendizado com um conjunto de dados usando um conjunto de treinamento, testar o seu desempenho usando um conjunto de validação, e tentar tirar conclusões a partir dos resultados obtidos. Os estatísticos definem uma metodologia para projetar corretamente experimentos e analisar os resultados obtidos de forma a ser capaz de extrair conclusões significativas. Baseado nessa metodologia, Alpaydin (2014) apresenta orientações para realização de experimentos de aprendizado de máquina que são apresentados e discutidos ao longo desta seção.

2.2.2.1 Fatores Controláveis e Incontroláveis

No aprendizado de máquinas, assim como em outros ramos da ciência e da engenharia, experimentos são realizados para obter informações do processo sob avaliação. No caso do aprendizado de máquina, o objeto da avaliação é um modelo de aprendizado criado a partir de um conjunto de dados de treinamento e que gera uma saída para uma determinada entrada (ALPAYDIN, 2014).

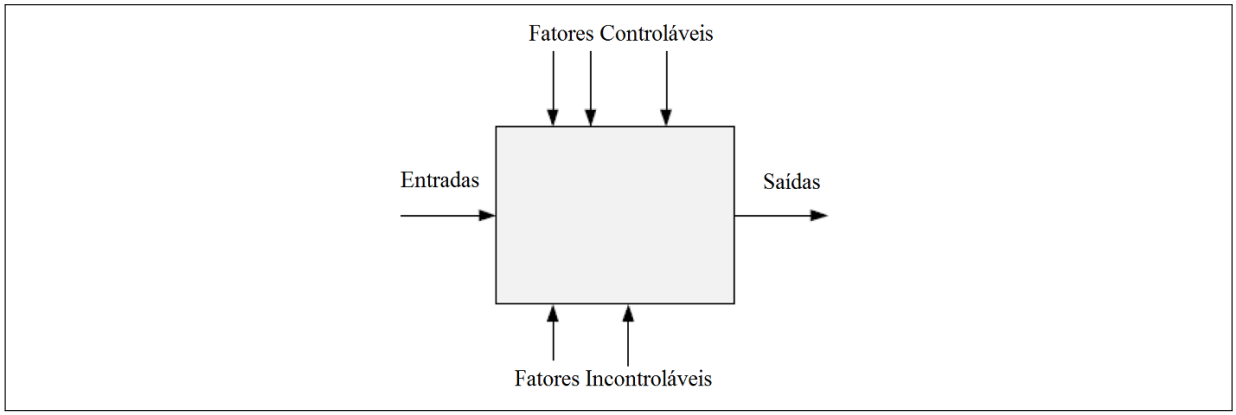
Um experimento é um teste ou uma série de testes em que são combinadas diferentes configurações dos fatores que afetam a saída do modelo de aprendizado. Esses fatores podem ser o algoritmo utilizado, o conjunto de treinamento, *features* de entrada, entre outros (ALPAYDIN, 2014). Nos testes, são combinadas diferentes configurações desses fatores e observadas as mudanças na saída, extraindo assim informações sobre a influência desses fatores sobre o modelo de aprendizado. O objetivo pode ser identificar os fatores mais importantes, os que não são importantes ou encontrar a configuração dos fatores que otimiza a saída. Ao planejar e conduzir experimentos de aprendizado de máquina, busca-se obter resultados estatisticamente significativos e conclusivos, eliminando ao máximo o efeito do acaso (ALPAYDIN, 2014).

Um modelo de aprendizado treinado é representado na Figura 5. Ele fornece uma saída que é dependente da entrada, dos fatores controláveis e dos fatores incontroláveis. Os fatores controláveis, como o nome sugere, são aqueles sobre os quais se tem controle ao planejar e executar um experimento (ALPAYDIN, 2014). Alguns exemplos dos fatores controláveis são os algoritmos de aprendizado utilizados, os hiperparâmetros dos algoritmos, os conjuntos de dados e as *features* de entrada extraídas dos dados.

Há também fatores incontroláveis que adicionam variabilidade indesejada ao pro-

cesso que podem afetar os resultados (ALPAYDIN, 2014). Alguns exemplos de fatores incontroláveis são os erros e exceções presentes nos dados e a aleatoriedade envolvida no processo, no algoritmo e na execução do experimento (ALPAYDIN, 2014).

Figura 5 – Fatores que influenciam experimentos de aprendizado de máquina.

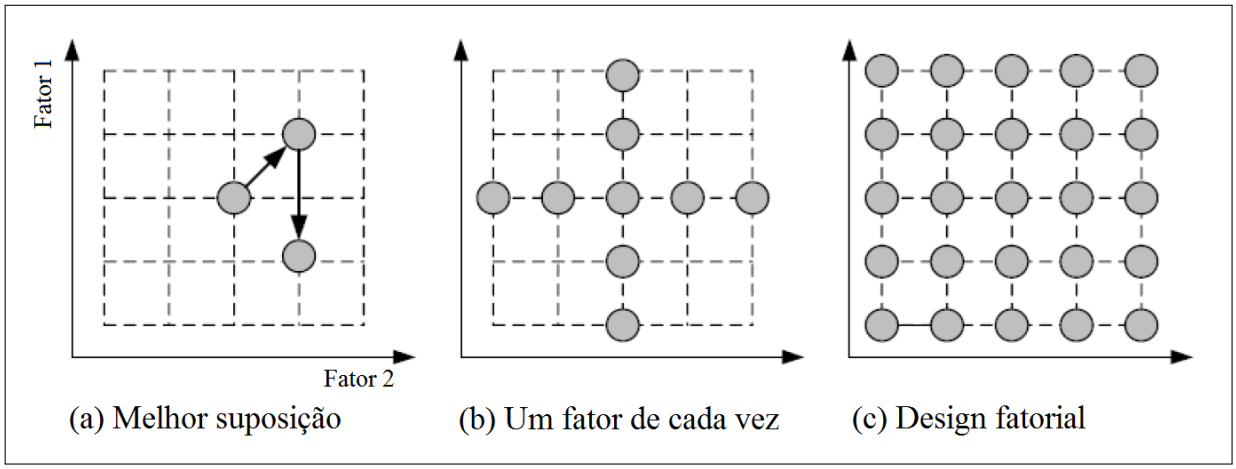


Fonte – O autor, traduzido de Alpaydin (2014).

2.2.2.2 *Estratégias de Experimentação*

De uma forma geral, é preciso determinar o efeito dos fatores controláveis sobre as saídas possíveis. No caso específico de problema de otimização, deve-se encontrar a configuração dos fatores que gera a melhor saída. Para isso, existem três estratégias possíveis para experimentos de aprendizado de máquina (ALPAYDIN, 2014) e elas são representadas na Figura 6 para o caso específico de dois fatores controláveis.

Figura 6 – Estratégias para experimentos de aprendizado de máquina.



Fonte – O autor, traduzido de Alpaydin (2014).

Na estratégia da melhor suposição, são testadas apenas poucas configurações supostamente boas até chegar a uma saída que seja considerada boa o suficiente (ALPAYDIN, 2014). Se o experimentador tem uma boa intuição do processo, isso pode funcionar. Entretanto, não há uma abordagem sistemática para modificar os fatores nem garantia de encontrar a melhor configuração.

Outra estratégia é alterar um fator de cada vez e manter todos os outros com um valor padrão, testando todas as configurações possíveis para o fator alterado (ALPAYDIN, 2014). A principal desvantagem dessa estratégia é que ela assume que não há relação entre os fatores, o que nem sempre é verdade.

A estratégia mais completa é o design fatorial. Nela todos os cenários possíveis são testados, considerando todas as configurações para os fatores avaliados (ALPAYDIN, 2014). Entretanto, dependendo do número de fatores e configurações possíveis, nem sempre essa estratégia é aplicável, tendo em vista a sua complexidade de tempo exponencial.

2.2.2.3 Medição de Desempenho do Classificador

Na classificação de problemas com duas classes, como é o caso do problema da autenticação, existem quatro medidas básicas possíveis (ALPAYDIN, 2014), descritas no Quadro 1. Para um exemplo positivo, se a predição é também positiva, este é um positivo verdadeiro (*True Positive* - TP); se a predição é negativa para um exemplo positivo, este é um falso negativo (*False Negative* - FN). Para um exemplo negativo, se a previsão também é negativa, este é um negativo verdadeiro (*True Negative* - TN); e, por fim, se a previsão é positiva para um exemplo negativo, este é um falso positivo (*False Positive* - FP). A partir dessas medidas básicas, as medidas de erro/acerto apresentadas no Quadro 2 podem ser calculadas e usadas para representar o desempenho estimado da aprendizagem para um problema específico (ALPAYDIN, 2014). Assim, a partir das saídas obtidas nos experimentos é possível produzir variáveis de resposta que permitem realizar a avaliação pretendida.

Em algumas aplicações de aprendizado de máquina, medir os recursos computacionais necessários é tão importante quanto medir os erros/acertos (ALPAYDIN, 2014). Nesses casos, outros critérios que podem ser considerados importantes envolvem a complexidade de tempo e espaço dos algoritmos (ALPAYDIN, 2014).

É relevante então mencionar que uma das características sabidamente comuns aos dispositivos móveis é a escassez de recursos. Por isso, neste trabalho considera-se importante

garantir o uso racional dos recursos utilizados nas aplicações de aprendizado de máquina em dispositivos móveis.

Quadro 1 – Matriz de confusão para algoritmos classificadores de duas classes

Classe	**Predição**		
Verdadeira	Positivo	Negativo	**Total**
Positivo	TP	FN	p
Negativo	FP	TN	n
Total	p'	n'	N

Fonte – O autor, traduzido de Alpaydin (2014)

Quadro 2 – Medidas de avaliação de desempenho para algoritmos de classificação

Medida	**Fórmula**
Erro	$(FP+FN)/N$
Acurácia	$(TP+TN)/N = 1-Erro$
TPR	TP/p
FPR	FP/n
AUC	Área sob a curva ROC
Precision	TP/p'
Recall	$TP/p = TPR$
F-Measure	$2*(Precision*Recall)/(Precision+Recall)$
Sensitivity	$TP/p = TPR$
Specificity	$TN/n = 1-FPR$

Fonte – O autor, traduzido de Alpaydin (2014)

2.2.2.4 Replicação e Validação Cruzada

Um dos princípios básicos do design experimental é a replicação (MONTGOMERY, 2012). Esse princípio determina que um experimento deve ser repetido um determinado número de vezes com uma mesma configuração dos fatores controláveis para que seja obtida uma média sobre o efeito dos fatores incontroláveis (MONTGOMERY, 2012). Como a saída varia nessas repetições do mesmo experimento, pode-se obter uma estimativa do erro experimental e determinar quando as diferenças entre os resultados obtidos com diferentes configurações dos fatores podem ser consideradas estatisticamente significativas.

No aprendizado de máquina, isto é comumente feito com a validação cruzada. Essas técnicas usam os mesmo dados divididos de formas diferentes para repetição do experimento (ALPAYDIN, 2014). Uma das técnicas de validação cruzada é a *K-Fold Cross-Validation* (ALPAYDIN, 2014). Essa técnica divide aleatoriamente os dados disponíveis em K partes de tamanhos iguais. Em seguida, são realizadas K repetições, sendo que em cada delas uma dessas K

partes é utilizada como conjunto de teste e as demais K - 1 partes como conjunto de treinamento. Para a análise dos resultados devem ser considerados a média obtida nas K repetições e o intervalo de confiança (ALPAYDIN, 2014). Os valores para K comumente utilizados são 10 ou 30 (ALPAYDIN, 2014).

2.2.2.5 Comparando Dois Algoritmos de Classificação

Outro princípio básico do design experimental é a blocagem (MONTGOMERY, 2012). Em experimentos de aprendizado de máquina, esse princípio estabelece que, por exemplo, os resultados obtidos com diferentes algoritmos devem variar apenas em função do algoritmo (ALPAYDIN, 2014). Assim, quando for usada a validação cruzada como técnica de replicação, é fundamental garantir também que todos os algoritmos sejam treinados e testados com os mesmos conjuntos de dados em cada um das K repetições, conforme apresentado na subseção 2.2.2.4. Desta forma, os resultados dos algoritmos não irão variar por conta de diferenças nos conjuntos de dados, o que prejudicaria os resultados obtidos.

Além disso, as comparações entre os resultados obtidos com os algoritmos diferentes devem utilizar testes de hipótese estatístico (ALPAYDIN, 2014). Assim, é possível verificar se esses resultados apresentaram variações estatisticamente significativas que possam representar evidência da diferença de desempenho entre os algoritmos avaliados. Uma das técnicas existentes para comparação de dois algoritmos de aprendizado é o Teste-T Pareado com Validação Cruzada K-Vezes (*K-Fold Cross-Validated Paired T-Test*) (ALPAYDIN, 2014). Com ela é possível calcular a diferença do resultado obtido por dois classificadores para um mesmo conjunto de treinamento/teste em cada uma das K repetições. Em seguida, essas K diferenças são usadas para testar estatisticamente se os desempenhos dos algoritmos podem ser considerados iguais ou se o desempenho de um deles pode ser considerado maior que o do outro.

2.2.2.6 Recursos Importantes

Existem alguns repositórios de conjuntos de dados na Internet que são usados frequentemente por pesquisadores de aprendizagem de máquina para fins de *benchmarking*. Ainda assim, alguns pesquisadores acreditam que tais repositórios não refletem as características completas de dados reais e são de escopo limitado (ALPAYDIN, 2014). Por isso, na avaliação de algoritmos de aprendizado, é recomendado ter o maior número possível de conjuntos de dados coletados a partir de aplicações semelhantes para realização de experimentos (ALPAYDIN,

2014).

Além disso, existem também pacotes de software livre que implementam vários algoritmos de aprendizagem de máquina. O WEKA (HALL *et al.*, 2009) é um deles e é especialmente notável (ALPAYDIN, 2014). Ele alcançou grande aceitação no meio acadêmico e tornou-se amplamente utilizado pelos pesquisadores, acumulando mais de 11 mil citações na biblioteca digital Google Scholar e 2 mil na ACM DL. O WEKA permite fácil acesso a um vasto conjunto de ferramentas usadas para aprendizado de máquina, incluindo pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização (HALL *et al.*, 2009). Outro ambiente de aprendizado de máquina disponível é o Microsoft Azure Machine Learning Studio[1]. Ele faz parte do conjunto de serviços de nuvem oferecidos na plataforma Microsoft Azure e tem foco na utilização profissional. Além desses ambientes citados, as linguagens Python[2] e R[3] também são muito populares entre os cientistas de dados, pois elas disponibilizam uma grande coleção de bibliotecas eficientes de aprendizado de máquina.

2.3 Considerações Finais

Neste capítulo foram apresentados conceitos e definições ligados à autenticação de usuários, com foco principalmente na biometria já que ela é fundamental para a autenticação transparente e contínua, foco desta dissertação.

Inicialmente foram apresentados de forma detalhada os principais requisitos que devem ser considerados para a utilização de uma traço biométrico, os processos que compõem um sistema biométrico e as medidas para avaliação do desempenho desses sistemas.

Além disso foram apresentados também conceitos e definições ligados ao aprendizado de máquina. O foco principal foi na importância da engenharia de *features* e no design de experimentos para avaliação de sistemas de aprendizagem. Nesse contexto, foram apresentados com detalhes fatores, estratégias, medidas, técnicas e recursos para a condução de experimentos necessários para a avaliação de desempenho desses sistemas.

Desta forma, com base neste capítulo é possível compreender a importância do aprendizado de máquina para a construção de sistemas biométricos, em especial voltados para a autenticação transparente e contínua em dispositivos móveis. Além disso, é possível compreender também a importância das *features* no desempenho tanto dos sistemas biométricos, como também

[1] https://studio.azureml.net
[2] https://www.python.org
[3] https://www.r-project.org/

das aplicações de aprendizado de máquina. Outro aspecto que este capítulo esclarece diz respeito à deficiência na avaliação de sistemas biométricos, além dos impactos disso na reprodutibilidade e comparação dos resultados das pesquisas publicadas na área e, consequentemente, na reutilização das técnicas propostas. Por fim, este capítulo permite compreender as boas práticas existentes no design de experimentos de aprendizado de máquina e a sua aplicabilidade na avaliação de sistemas biométricos.

Com base nisso e nos trabalhos relacionados que são apresentados no Capítulo 3, foi elaborado um processo de avaliação das *features* biométricas que é descrito em detalhes no Capítulo 4.

3 TRABALHOS RELACIONADOS

Neste capítulo são apresentados os trabalhos relacionados. As seções deste capítulo estão estruturadas da seguinte forma. A seção 3.1 apresenta os trabalhos relacionados com foco na padronização de métodos para a avaliação biométrica, destacando as suas principais características e também as principais semelhanças e diferenças com o objetivo desta pesquisa. Na sequência, a seção 3.2 apresenta os trabalhos relacionados que incluem a utilização de *features* de localização *outdoor* para a autenticação transparente e contínua em dispositivos móveis, utilizadas no estudo de caso apresentado no Capítulo 5. A seção 3.3 apresenta uma discussão sobre os trabalhos relacionados identificados. Por fim, a seção 3.4 apresenta as considerações finais.

3.1 Avaliação Biométrica

Durante a busca por trabalhos relacionados focados na padronização de métodos para a avaliação biométrica foram encontradas referências sobre o projeto BEAT[1] (*Biometrics Evaluation and Testing*) (MARCEL, 2013). Esse projeto é parte do programa específico FP7-SECURITY[2] financiado pela Fundo de Pesquisa e Inovação da Comissão Europeia.

O projeto BEAT (MARCEL, 2013) foi concluído em 2013 e teve como objetivo propor um *framework* para avaliação operacional padronizada de tecnologias biométricas. Para alcançar esse objetivo, esse projeto incluiu: o desenvolvimento de uma plataforma online[3] e aberta para avaliar de forma transparente e independente os sistemas biométricos usando bases de dados de referência e validadas; o projeto de protocolos e ferramentas para análise de vulnerabilidade; e a elaboração de documentos de padronização para avaliação da segurança com base no padrão internacional ISO/IEC-15408-1 (2009). Além disso, o projeto BEAT (MARCEL, 2013) considerou aspectos legais e abordou as questões de proteção de dados, privacidade e propriedade intelectual, buscando garantir assim que o *framework* BEAT pode ser usado tanto pela comunidade de pesquisa, como também pelas empresas.

Para o desenvolvimento da sua plataforma de avaliação online e aberta, o projeto BEAT desenvolveu uma ferramenta de código aberto para processamento de sinal e aprendizado de máquina denominada BOB (ANJOS *et al.*, 2012). Segundo os seus desenvolvedores, o BOB

[1] http://cordis.europa.eu/project/rcn/102363_en.html
[2] http://cordis.europa.eu/programme/rcn/861_en.html
[3] https://www.beat-eu.org/platform/

é uma ferramenta projetada para atender às necessidades dos pesquisadores que reduz o tempo de desenvolvimento e garante o processamento eficiente dos dados. O desenvolvimento rápido é possível por meio de um ambiente Python amigável disponibilizado para o pesquisador. Já o processamento eficiente de grandes quantidades de dados é fornecido por implementações em C++ dos gargalos identificados na ferramenta. O ambiente Python é integrado com as bibliotecas C++ e garante a sua extensividade. Além disso, a ferramenta BOB (ANJOS *et al.*, 2012) apoia a reprodutibilidade da pesquisa através de seus protocolos integrados para realização de experimentos com várias bases de dados. Portanto, ela pode ser um recurso útil para os pesquisadores devido a sua combinação de facilidade de uso, eficiência, extensibilidade e transparência.

Assim como o projeto BEAT, esta pesquisa tem foco na padronização de métodos para a avaliação biométrica, reprodutibilidade da pesquisa e reuso de técnicas. No entanto, esta pesquisa não tem foco na avaliação de um sistema biométrico como um todo, mas apenas das *features* biométricas utilizadas ou que se deseja utilizar para a construção desses sistemas. Por isso, o objetivo deste trabalho é propor um processo para a avaliação de *features* biométricas usando aprendizado de máquina e realizar um estudo de caso com esse processo usando *features* de localização.

De fato, os sistemas biométricos são basicamente sistemas de reconhecimento de padrões (JAIN *et al.*, 2007)(ALPAYDIN, 2014) e é notório que as técnicas de aprendizado de máquina são largamente utilizadas para construção desses sistemas biométricos (PATEL *et al.*, 2016)(KHAN, 2016)(ANJOS *et al.*, 2012). Além disso, como já foi apresentado no Capítulo 2, as técnicas de aprendizado de máquina são de propósito geral e podem ser reutilizadas com facilidade, pois já foram bem dominadas (DOMINGOS, 2012)(ALPAYDIN, 2014). Por isso, a parte mais importante para o desenvolvimento desses sistemas é a engenharia de *features* (DOMINGOS, 2012). Assim, para este trabalho, a chave para a melhoria do desempenho dos sistemas biométricos são as *features* biométricas.

Mesmo com a diferença de foco, foi verificada a possibilidade de utilização da plataforma online e aberta disponibilizada pelo projeto BEAT (MARCEL, 2013) para a realização dos experimentos que fazem parte do estudo de caso para avaliação das *features* de localização identificadas nos trabalhos relacionados que são apresentados na seção 3.2 deste capítulo. No entanto, ao analisar de forma mais aprofundada os recursos disponibilizados por essa plataforma foi possível perceber que atualmente o seu foco se concentra na biometria baseada na face,

impressão digital, íris e padrões de vasos sanguíneos. Por isso, embora a plataforma seja extensível, ela não foi utilizada. No lugar dela, foi utilizada a ferramenta de aprendizado de máquina WEKA (HALL *et al.*, 2009) que é amplamente utilizada em pesquisas científicas, como já foi apresentado na subseção 2.2.2.6. Ela foi escolhida porque disponibiliza nativamente todos os algoritmos necessários para a realização deste trabalho.

3.2 *Features* de Localização

Na fase exploratória desta pesquisa foram identificados dois estudos (PATEL *et al.*, 2016)(KHAN, 2016) que apresentam amplas revisões da literatura com trabalhos que utilizam *features* extraídas de diferentes traços biométricos para construção de mecanismos de autenticação transparente e contínua em dispositivos móveis. Analisando esses estudos foi possível identificar que Patel *et al.* (2016) apresenta dois trabalhos que utilizam *features* de localização *outdoor* (SHI *et al.*, 2011)(FRIDMAN *et al.*, 2016) e Khan (2016) apresenta outros três trabalhos com essa mesma *feature* (JAKOBSSON *et al.*, 2009)(BUTHPITIYA *et al.*, 2014)(KAYACIK *et al.*, 2014).

Na tentativa de identificar mais trabalhos que também utilizam *features* de localização *outdoor*, esta pesquisa realizou uma revisão da literatura seguindo uma parte do protocolo sistemático de busca proposto por Kitchenham e Charters (2007). Esta revisão incluiu uma busca nos metadados de três bases científicas (*Science Direct*, ACM DL e IEEE Xplore DL) por trabalhos publicados nos últimos dez anos usando a string de busca: *(authentication OR authenticate) AND (implicit OR transparent OR continuous) AND (mobile OR smartphone) AND (location OR gps OR spacial OR spatial OR spatio)*.

Nessa busca trinta e seis artigos foram identificados, sendo três da *Science Direct*, treze da ACM DL e vinte da IEEE Xplore DL. Depois de duas rodadas de análise, observando os critérios de inclusão e exclusão definidos para essa revisão com base no escopo desta pesquisa, restaram três artigos, sendo dois da ACM DL (RAMAKRISHNAN *et al.*, 2015)(LIMA *et al.*, 2011) e um da IEEE Xplore DL (TANG *et al.*, 2010).

Além desses, mais um trabalho (HAYASHI *et al.*, 2013) foi identificado usando as palavras-chave da string de busca na ferramenta *Google Scholar*.

Os principais aspectos analisados nesses nove trabalhos identificados são apresentadas de forma resumida no Quadro 3 e descritos em detalhes ao longo do restante desta seção.

No primeiro trabalho, Jakobsson *et al.* (2009) analisaram o uso de dados de ligação

Quadro 3 – Aspectos analisadas nos trabalhos relacionados.

Trabalho	**Tecnologia**	***Features***	**Medidas**	**Cj. de Dados**
Jakobsson *et al.* (2009)	GPS	Latitude e Longitude	Outras	Coletado pelo autor
Tang *et al.* (2010)	GPS	Latitude, Longitude e Tempo	TPR e FPR	Coletado pelo autor (10 usuários)
Lima *et al.* (2011)	GPS	Latitude, Longitude e Tempo	Outras	Coletado pelo autor (280 eventos)
Shi *et al.* (2011)	Celular	Sequência de Células	TPR, FPR, *Precision* e *Recall*	Coletado pelo autor (7 usuários)
Hayashi *et al.* (2013)	GPS	Latitude e Longitude	Outras	Coletado pelo autor (32 usuários)
Buthpitiya *et al.* (2014)	Celular	Sequência de Células e Tempo	TPR, FPR, curva ROC, Acurácia e Energia	Coletado pelo autor (30 usuários)
Kayacik *et al.* (2014)	Celular	Célula e Tempo	Outras	Público (132 usuários)
Ramakrishnan *et al.* (2015)	GPS	Latitude e Longitude	TPR e FPR	Coletado pelo autor (4 usuários)
Fridman *et al.* (2016)	GPS	Latitude e Longitude	TPR, FPR e curva ROC	Coletado pelo autor (200 usuários)

Fonte – O autor.

realizada e localização como traços biométricos para autenticação transparente e contínua. Nesse trabalho os dados de latitude e longitude obtidos por meio de GPS são usados como *features* de localização. Já o aprendizado acontece por meio de clusterização com o algoritmo DJ-Cluster (ZHOU *et al.*, 2007) para aprender os locais mais visitados. Esse algoritmo é parametrizado pelo número mínimo de pontos e a distância máxima permitida entre esses pontos dentro de um mesmo agrupamento. Segundo os autores, com os valores adequados, o algoritmo produz um pequeno número de agrupamentos correspondentes ao local onde o usuário vive, trabalha e realiza compras. Nesse trabalho não foi realizada uma avaliação da proposta usando medidas. Em vez disso, foi realizada apenas uma análise preliminar usando dados coletados durante a pesquisa a partir de um número não especificado de indivíduos usando dispositivos móveis BlackBerry por um período de 3 meses.

No segundo, Tang *et al.* (2010) apresentam um esquema de autenticação de usuários móveis usando um método de mineração de dados que realiza a identificação com base na aplicação usada e a localização. Nesse trabalho a localização é obtida por GPS e as *features* de localização utilizadas são latitude, longitude e tempo. Os dados são organizados em gráficos direcionais e os usuários são identificados por meio de um classificador baseado em regras. Estes dados precisam ser pré-processados para que as localizações pertencentes a uma mesma área quadrada formem um grupo. Na avaliação, os autores usaram dados coletados durante a pesquisa

a partir de 10 voluntários usando dispositivos móveis iPhone por um período de 20 dias. As medidas usadas na avaliação foram TPR e FPR.

No terceiro, Lima *et al.* (2011) apresentam o CARS-AD, uma arquitetura de autenticação sensível ao contexto que usa dados de ligação realizada, atividade agendada, aplicação usada, uso de bateria e localização como traços biométricos. Baseado nesses elementos, é possível compor um modelo probabilístico que identifica situações de uso normal, anormal ou suspeito do dispositivo. Nesse trabalho, assim como em Fridman *et al.* (2016), os dados de latitude e longitude obtidos por meio de GPS são usados como *features* de localização. No entanto, nesse trabalho é usado também como *feature* o instante de tempo em que a localização foi visitada. Na avaliação, os autores usaram dados de 280 eventos coletados durante a pesquisa. As medidas usadas na avaliação mensuram o grau de similaridade entre os perfis obtidos versus o número de interações do usuário.

No quarto, Shi *et al.* (2011) apresentam o SenGuard, um *framework* para autenticação contínua e implícita de usuários para *smartphone*. O protótipo implementado desse *framework* utilizou dados de toque na tela, acelerômetro, voz e localização para a extração de *features* biométricas. Nesse trabalho, os dados de localização foram obtidos por meio do Sistema de Telefonia Celular e as *features* de localização extraídas foram formadas por sequências de identificadores de células que representam torres de telefonia com as quais o dispositivo estabeleceu conexão enquanto realiza um deslocamento. Os autores desse trabalho defendem que uma célula representa a localização do usuário em um nível grosseiro, mas esse método de localização é mais eficiente energeticamente quando comparado com outros métodos de localização como o GPS e a célula pode ser obtida em quase qualquer dispositivo celular, hora e local. No entanto, nenhuma medida é utilizada para avaliar o consumo de energia nesse trabalho. O SenGuard aprende sequências de células usando uma janela deslizante de tamanho fixo 6 que representa o número de células incluídas em cada sequência. Durante o aprendizado, as sequências coletadas na fase de treinamento são transformadas em padrões de trajetória do usuário do *smartphone*. Em seguida, as sequências coletadas depois da fase de treinamento são comparadas com os padrões aprendidos, medindo a distância de Levenshtein entre elas. Essa distância representa um desvio do trajeto aprendido. Se a distância medida for maior que um limite definido, o algoritmo indica que o usuário atual não é o proprietário do *smartphone*. Na avaliação, os autores utilizaram dados coletados a partir de 7 indivíduos durante a pesquisa. As medidas usadas na avaliação foram TPR, FPR, *Precision* e *Recall*.

No quinto, Hayashi *et al.* (2013) apresentam um *framework* probabilístico para a seleção dinâmica de um esquema de autenticação ativa (*e.g.* digitar uma senha) a partir de requisitos de segurança especificados pela aplicação e traços biométricos que podem ser coletados de forma transparente e contínua do usuário. Um protótipo desse *framework* foi desenvolvido usando apenas a localização como traço biométrico passivo. Nesse trabalho, assim como em Jakobsson *et al.* (2009), os dados de latitude e longitude obtidos por meio de GPS são usados como *features* de localização. Essas *features* são agrupadas com o algoritmo de aprendizado *K-Nearest Neighbor* (K-NN) e associados com as localizações pré-definidas pelo sistema (*i.e.* casa, trabalho e outras). Na avaliação, os autores usaram dados coletados durante a pesquisa a partir de 32 indivíduos usando dispositivos móveis Android ao longo de pelo menos 7 dias. As medidas usadas na avaliação foram os lugares onde os usuários gastam mais tempo, o número de ativações do dispositivo feitas nesses lugares e a taxa de acerto na identificação da localização atual durante o processo de autenticação.

No sexto, Buthpitiya *et al.* (2014) avaliaram o uso de *features* biométricas extraídas de mensagem enviada, ligação realizada, localização *outdoor* e *indoor* para autenticação transparente e contínua. Nesse trabalho, assim como em Shi *et al.* (2011), os dados de localização foram obtidos por meio do Sistema de Telefonia Celular e as *features* de localização extraídas foram formadas por sequências de identificadores de células que representam torres de telefonia com as quais o dispositivo estabeleceu conexão enquanto realizava um deslocamento. No entanto, o instante de tempo em que o dispositivo estabelece a conexão com uma torre celular é utilizado como *feature* nesse trabalho. Além disso, também como em Shi *et al.* (2011), os autores desse trabalho defendem que uma célula representa a localização do usuário em um nível grosseiro, mas esse método de localização é mais eficiente energeticamente quando comparado com outros métodos de localização como o GPS. No entanto, nesse trabalho foram incluídas medidas para avaliar o consumo de energia. Já o aprendizado é realizado por meio do cálculo da Estimativa da Máxima Probabilidade (*Maximum Likelihood Estimation* - MLE) de uma sequência de células ser visitada em um determinado instante de tempo. Na avaliação, os autores usaram dados coletados durante a pesquisa a partir de 30 indivíduos usando dispositivos móveis Android por um período de 3 meses. As medidas usadas na avaliação foram TPR, FPR, Acurácia, curva ROC e consumo de energia da bateria.

No sétimo, Kayacik *et al.* (2014) avaliaram o uso de *features* extraídas de várias fontes de dados dos dispositivo, incluindo a localização, para autenticação transparente e contínua.

Nesse trabalho, assim como em Buthpitiya *et al.* (2014), os dados de localização são obtidos por meio do Sistema de Telefonia Celular e envolvem identificadores de célula e o instante de tempo da conexão com torre celular. No entanto, nesse trabalho não são usadas sequências de células como *features*. Já o aprendizado se dá por meio da Função de Densidade de Probabilidade (*Probability Density Function* - PDF) construída a partir de histogramas que contabilizam as ocorrências de cada localização por instante de tempo. Na avaliação, os autores usam 3 (três) conjuntos de dados disponíveis publicamente que somam juntos um total de 132 indivíduos usando dispositivos móveis por um período que variou de poucos dias até próximo de 1 ano, dependendo do usuário. As medidas usadas na avaliação estão ligadas ao conforto do usuário e detecção de ataques.

No oitavo, Ramakrishnan *et al.* (2015) descrevem o PRISM, um *framework* que fornece autenticação implícita baseada em risco através da detecção de anomalias nos padrões de comportamento habituais dos usuários. Esse *framework* utiliza dados de atividade agendada, aplicação usada e localização *indoor* e *outdoor* como traços biométricos. Assim como em Jakobsson *et al.* (2009), a localização *outdoor* é obtida por GPS e as *features* são formadas pela latitude e longitude. Nesse trabalho, as decisões de desbloqueio do dispositivo são dirigidas por políticas que são definidas automaticamente pela mineração de dados de sensores ou manualmente pelos usuários finais. As *features* são agrupadas com o algoritmo de aprendizado *K-Nearest Neighbor* (K-NN) e associadas com as localizações definidas pelo usuário (*e.g.* casa, trabalho e outras). Na avaliação, os autores usaram dados coletados durante a pesquisa a partir de 4 indivíduos usando dispositivos móveis Android por um período que variou de 3 a 13 dias. As medidas usadas na avaliação foram TPR e FPR.

No nono e último trabalho, Fridman *et al.* (2016) avaliaram o uso de dados de estilo linguístico, aplicação usada, sítio web acessado e localização como traços biométricos para autenticação transparente e contínua. Nesse trabalho, os dados de latitude e longitude obtidos por meio de GPS são usados como *features* de localização. Já o aprendizado acontece por meio de classificação com o algoritmo *Support Vector Machine* (SVM) e o kernel *Radial Basis Function* (RBF). Com base no modelo treinado é possível calcular a probabilidade de uma localização fazer parte do padrão de deslocamento do usuário. Na avaliação, os autores usaram dados coletados durante a pesquisa a partir de 200 indivíduos usando dispositivos móveis Android por um período de pelo menos 30 dias. As medidas usadas na avaliação foram TPR, FPR e curva ROC.

Por fim, a partir desses trabalhos foram elaborados quatro conjuntos de *features* de localização que representam a combinação das diferentes estratégias encontradas na literatura. De forma geral, esses conjuntos de *features* buscam representar no estudo de caso realizado nesta pesquisa as estratégias que utilizam como *features* uma localização isolada ou sequencias de localizações, combinadas ou não com o instante de tempo em que essas localizações foram registradas. No Capítulo 5 são apresentados mais detalhes sobre as *features* elaboradas a partir dos trabalhos relacionados apresentados nessa seção e também sobre o estudo de caso para a avaliação do processo proposto nesta pesquisa.

3.3 Discussão

A partir da revisão da literatura realizada e apresentada neste capítulo foi possível constatar que a avaliação biométrica não têm sido um tema muito atacado por outros pesquisadores. No caso específico da avaliação de *features* biométricas, não foi possível encontrar nenhum outro trabalho relacionado ao tema. Por outro lado, foi possível constatar que essa é uma questão importante ao analisar de forma específica e aprofundada os trabalhos na área de autenticação transparente e contínua em dispositivos móveis que fazem uso de *features* de localização. Os trabalhos identificados nessa área apresentam avaliações que não facilitam a comparação dos resultados com outras pesquisas. Isso não permite que a comunidade acadêmica possa confirmar com clareza as contribuições trazidas por essas propostas ou mesmo determinar o estado da arte atual. Além disso, os trabalhos também não descrevem com detalhes as técnicas e procedimentos utilizados. Isso dificulta a tentativa de reprodução dos experimentos realizados. Por fim, quase todos os trabalhos identificados, com exceção apenas de Kayacik *et al.* (2014), não usam nas suas avaliações dados disponíveis publicamente, mesmo quando eles existem e são disponibilizados com facilidade. Isso impede a reprodução dos resultados obtidos com os experimentos.

Sendo assim, com a análise da literatura nesta dissertação, é possível identificar a necessidade de um processo para avaliação de *features* biométricas para autenticação transparente e contínua em dispositivos móveis baseado em aprendizado de máquina. Este processo deve trazer detalhes das técnicas e procedimentos a serem utilizados neste tipo de avaliação para garantir a reprodutibilidade dos experimentos e a comparação dos resultados com outras pesquisas. A proposta deste processo é o objetivo principal desta dissertação.

A partir da análise desses trabalhos relacionados também é possível definir conjuntos que representam as variações das *features* de localização *outdoor* utilizadas pelas propostas de

autenticação transparente e contínua para dispositivos móveis encontradas na literatura. Além disso, é possível determinar também os algoritmos de aprendizado de máquina comumente utilizados para o reconhecimento de padrões de localização nessa mesma área de pesquisa. Esses dois fatores representam aspectos importantes e que podem ser usados para a realização de um estudo de caso. O resultado obtido a partir da realização de um estudo de caso é uma das contribuições previstas nesta dissertação e é o meio pelo qual busca-se avaliar o processo que representa o principal objetivo desta pesquisa.

3.4 Considerações Finais

Este capítulo concentrou-se mais nos trabalhos relacionados à avaliação biométrica e às *features* de localização, visto que não foram encontradas pesquisas relacionadas ao processo de avaliação biométrica propriamente dito, embora seja evidente a necessidade do mesmo, conforme explicitado no decorrer deste capítulo.

Os detalhes do processo proposto são descritos no Capítulo 4. Já os detalhes da realização do estudo de caso são apresentados no Capítulo 5.

4 PROCESSO DE AVALIAÇÃO DE *FEATURES* BIOMÉTRICAS

Neste capítulo é apresentado o processo proposto neste trabalho para avaliação de *features* biométricas para autenticação transparente e contínua em dispositivos móveis baseado em aprendizado de máquina. Além disso, são apresentadas também as justificativas para as principais decisões tomadas para a elaboração desse processo com base nas referências utilizadas.

Este processo permite a avaliação por meio de experimentos e a obtenção de resultados conclusivos sobre o desempenho das *features* biométricas utilizadas para a construção de sistemas biométricos. Essa avaliação é especialmente útil durante o processo de tentativa e erro que compreende a engenharia de *features* descrita por Domingos (2012). Além disso, ela é útil também para a comparação do desempenho de novas *features* com outras equivalentes identificadas em trabalhos relacionados.

As seções deste capítulo estão estruturadas da seguinte forma: a seção 4.1 apresenta as principais referências utilizadas para a elaboração do processo proposto; na sequência, a seção 4.2 descreve de forma geral o processo elaborado; as seções 4.3 a 4.6 apresentam os detalhes dos conjuntos de atividades que compõem esse processo; e, por fim, a seção 4.7 apresenta as considerações finais.

4.1 Introdução

Segundo Jain (1990), a maioria dos problemas de avaliação de desempenho consiste basicamente em encontrar o melhor entre um conjunto de alternativas. Esse é exatamente o problema encontrado quando se tenta avaliar *features* biométricas. Isso ocorre durante o desenvolvimento de um novo sistema biométrico, ou mesmo quando simplesmente busca-se comparar *features* diferentes de um mesmo traço biométrico utilizadas por diferentes sistemas.

Além disso, Jain (1990) identificou vários erros frequentemente cometidos na avaliação de sistemas computacionais. Por exemplo, se uma medição é repetida várias vezes, geralmente os resultados são ligeiramente diferentes em cada uma delas. Isso é comum e tem relação com os fatores incontroláveis envolvidos nos experimentos, conforme apresentado com mais detalhes na subseção 2.2.2.1. Assim, simplesmente comparar a média dos resultados obtidos em uma série de ensaios repetidos não leva a conclusões corretas, especialmente quando a variabilidade desse resultado é alta. Baseado nisso, ele propôs um conjunto de recomendações que buscam evitar esses erros comuns. Uma dessas recomendações diz respeito à condução

de avaliações de desempenho utilizando técnicas estatísticas adequadas. Essas recomendações são usadas em vários contextos de avaliação de sistemas computacionais, como por exemplo na avaliação de desempenho dos mecanismos de segurança usados em redes de sensores sem fio (CAVALCANTE, 2012)

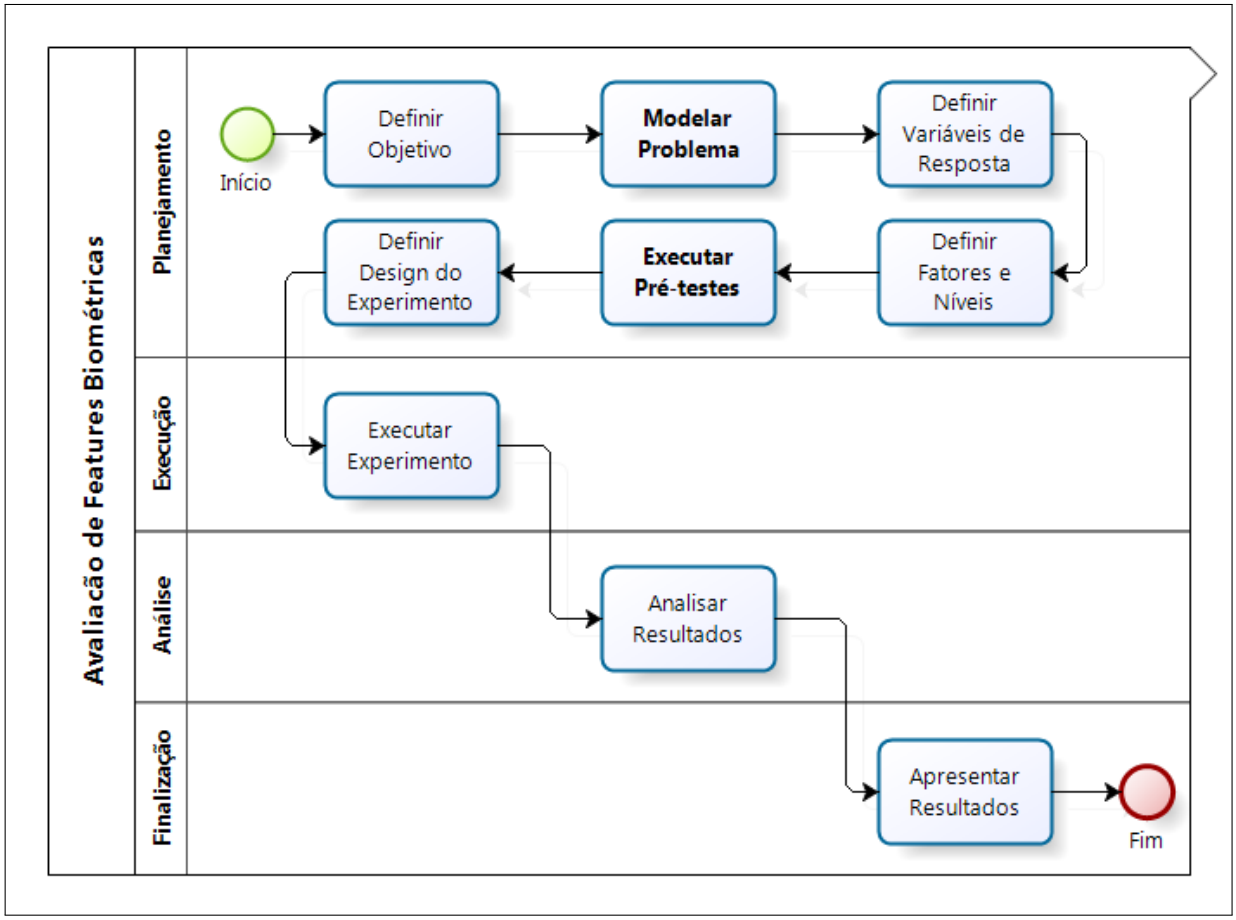
Como citado na seção 1.2, ao longo desta pesquisa foram identificadas na literatura deficiências nas técnicas estatísticas utilizadas na condução de experimentos, realizados para avaliação de propostas de mecanismos de autenticação transparente e contínua. Além disso, os únicos trabalhos com foco nessa problemática na área da biometria encontrados ao longo desta pesquisa foram apresentados na seção 3.1. Por outro lado, como foi discutido no Capítulo 2, as técnicas de aprendizado de máquina podem ser utilizadas para a construção de sistemas biométricos e as técnicas de avaliação utilizadas nessas duas áreas guardam muitas semelhanças, já que um sistema biométrico de fato é um sistema de reconhecimento de padrões. Ademais, alguns trabalhos evidenciam o avanço percebido nas últimas décadas nas avaliações realizadas nas pesquisas na área de aprendizado de máquina (BOUCKAERT; FRANK, 2004)(DEMŠAR, 2006)(DIETTERICH, 1998). Por isso, nesta dissertação busca-se suprir essas deficiências identificadas na avaliação de propostas de mecanismos de autenticação transparente e contínua usando aprendizado de máquina.

Além das recomendações propostas por Jain (1990), o processo proposto nesta dissertação considerou também as recomendações de Alpaydin (2014) apresentadas no Capítulo 2 deste trabalho, que têm como objetivo o planejamento e a condução de experimentos de aprendizado de máquina com base em uma metodologia estatística (MONTGOMERY, 2012). As recomendações de Alpaydin (2014) foram escolhidas para embasar esse processo porque representam uma das referência mais atualizadas e utilizadas na área de aprendizado de máquina. Além disso, o autor inclui um capítulo completo dedicado ao design e análise de experimentos com foco nas técnicas utilizadas em aprendizado de máquina, essencial para este trabalho.

4.2 O Processo

Para a modelagem do processo proposto neste trabalho foi utilizada a notação BPMN (CBOK, 2013). O processo modelado, apresentado na Figura 7, é composto por atividades e especialidades previstas para a avaliação de *features* biométricas para autenticação transparente e contínua em dispositivos móveis e é baseado em aprendizado de máquina. Esse processo inclui nove atividades que foram definidas com base nas recomendações de Alpaydin (2014)

Figura 7 – Processo proposto para a avaliação de *features* biométricas.



Fonte – O autor.

e Jain (1990). No entanto, as atividades em negrito não representam passos explicitamente recomendados por esses autores, mas foram incluídas no processo para atender as necessidades específicas da avaliação de *features* biométricas usando aprendizado de máquina.

As atividades foram divididas em quatro especialidades: Planejamento, Execução, Análise e Finalização. Essas especialidades representam os papéis ou perfis dos profissionais envolvidas com os respectivos conjuntos de atividades do processo. Esse agrupamento das atividades é importante, pois projetos reais de avaliação de desempenho podem ser grandes e exigir a coordenação de vários profissionais com diferentes habilidades (JAIN, 1990).

Os detalhes deste processo são discutidos nas próximas seções deste capítulo, onde são apresentados o objetivo, a descrição, as entradas e as saídas de cada atividade por especialidade, juntamente com as justificativas baseadas nas referências utilizadas.

4.3 Planejamento

Para Jain (1990), o planejamento de uma avaliação deve ser realizado por analistas de desempenho experientes e com conhecimento profundo sobre o problema a ser avaliado. Segundo ele, analistas inexperientes frequentemente:

- Subestimam a importância do planejamento e começam os projetos pela codificação e execução dos experimentos, pois sentem que nada realmente foi conseguido até que alguns resultados numéricos tenham sido obtidos;
- Adotam abordagens não sistemáticas, o que os leva a selecionar arbitrariamente variáveis de resposta, fatores e níveis, gerando conclusões imprecisas; e
- Costumam escolher medidas que podem ser facilmente calculadas ou coletadas, em vez daquelas que são relevantes para o objetivo da avaliação.

Ainda segundo Jain (1990), com a experiência, os analistas aprendem que uma grande parte do esforço de análise de um projeto de avaliação é aplicado nas atividades de planejamento. Apenas a compreensão do problema e a definição do objetivo geralmente levam até 40% do esforço total do projeto. Dos 60% restantes, grande parte ainda é dedicada à definição dos fatores e níveis, e também do design do experimento.

No restante desta seção serão apresentados os detalhes das atividades do processo que são de responsabilidade dos profissionais de planejamento.

4.3.1 Definir Objetivo

Toda avaliação de desempenho deve ter início com a definição de um objetivo (ALPAYDIN, 2014)(JAIN, 1990). A escolha de medidas corretas, conjuntos de dados representativos e metodologia apropriada depende do objetivo definido para a avaliação (JAIN, 1990). A necessidade de um objetivo pode parecer óbvia, mas a ausência de objetivos claros, a definição de objetivos tendenciosos e a análise sem a compreensão do problema são erros comuns nessa área (JAIN, 1990).

Em aprendizado de máquina, uma avaliação pode ser realizada para identificar qual entre dois algoritmos de aprendizagem apresenta menor erro esperado usando um conjunto de dados representativo (ALPAYDIN, 2014). Esses algoritmos podem ser completamente diferentes ou, por exemplo, usar apenas técnicas de extração de *features* diferentes (ALPAYDIN, 2014).

Nessa atividade deve ser definido o que se pretende alcançar com a avaliação pre-

tendida. Para isso, é importante a compreensão daquilo que será avaliado e o problema a ser resolvido (JAIN, 1990). No entanto, isso não é trivial, pois a maioria dos problemas de desempenho são vagos quando apresentados pela primeira vez e entender o problema o suficiente para definir o objetivo é complexo (JAIN, 1990). Com base nisso, o Quadro 4 apresenta os detalhes da atividade Definir Objetivo do processo proposto.

Quadro 4 – Detalhes da atividade Definir Objetivo.

Objetivo: Definir o que é desejado alcançar com a avaliação.	
Descrição: Nesta atividade deve ser compreendido o problema e depois definido o que se pretende alcançar com a avaliação. Para isso, no caso da avaliação de *features* biométricas, devem ser consideradas as opções de *features* biométricas disponíveis e o contexto de aplicação. Os principais aspectos relacionados ao contexto de aplicação que devem ser considerados são os requisitos que o sistema busca atender, os problemas esperados e como as *features* biométricas podem contribuir.	
Entradas: • *Features* biométricas • Informações sobre o contexto de aplicação	**Saídas:** • Objetivo da avaliação

Fonte – O autor.

4.3.2 Modelar Problema

A atividade Modelar Problema não é baseada nas recomendações de Jain (1990) e Alpaydin (2014), mas é necessária no caso da avaliação de *features* biométricas usando aprendizado de máquina. Nessa atividade, um problema de aprendizado de máquina deve ser modelado para que um problema de autenticação possa ser resolvido usando as *features* biométricas. Por exemplo, no passo Modelar Problema devem ser feitas escolhas sobre o tipo de aprendizado (Supervisionado ou Não Supervisionado), de algoritmo (Classificação, Regressão ou Clusterização), de algoritmo de classificação (One-class, Two-class ou Multiclass), entre outras.

Assim como na definição do objetivo, na modelagem do problema é importante a compreensão daquilo que será avaliado e o problema a ser resolvido. Essa modelagem deve ser compatível com um modelo de segurança e de ameaças idealizado para o sistema onde se pretende fazer uso das *features* biométricas avaliadas. Por exemplo, esse modelo pode prever a verificação da identidade de um único usuário autorizado de um dispositivo com base apenas no seu padrão comportamental. Outro exemplo de modelo pode prever essa verificação para múltiplos usuários autorizados com base no padrão comportamental de cada um deles. No segundo exemplo pode ser possível a modelagem usando algoritmos de classificação two-class

ou multiclass, já que existem dados de vários usuários disponíveis no contexto de aplicação e cada usuário pode representar uma classe em cada classificador treinado. No entanto, no primeiro exemplo pode ser possível apenas uma modelagem one-class, já que no contexto de aplicação só existem dados de um usuário para treinamento.

A modelagem do problema é necessária para o processo proposto porque é preciso transformar um problema de autenticação em um problema de aprendizado. Assim, o problema de autenticação passa a ser compatível com as diretrizes propostas por Alpaydin (2014) para realização de experimentos de aprendizado de máquina. Essa modelagem deve ser feita imediatamente depois da definição do objetivo, pois a modelagem influencia todas as outras atividades do processo. A modelagem permitirá a definição das variáveis de resposta adequadas e dos fatores determinantes para o desempenho das *features* biométricas avaliadas. Com base nisso, o Quadro 5 apresenta os detalhes da atividade Modelar Problema do processo proposto.

Quadro 5 – Detalhes da atividade Modelar Problema.

<table>
<tr><td colspan="2">Objetivo: Modelar um problema de aprendizado de máquina que represente satisfatoriamente o problema de autenticação no qual as features biométricas serão aplicadas.</td></tr>
<tr><td colspan="2">Descrição: Nesta atividade deve ser compreendido o modelo de segurança e ameaças previsto para o problema de autenticação onde as features biométricas serão utilizadas. Além disso, deve ser modelado um problema de aprendizado de máquina que represente satisfatoriamente esse problema de autenticação. Assim como na definição do objetivo, as opções de features biométricas disponíveis e o contexto de aplicação devem ser consideradas na modelagem do problema. O principal aspecto relacionado ao contexto de aplicação que deve ser considerado é o modelo de segurança e ameaças previsto para o problema de autenticação.</td></tr>
<tr><td>Entradas:
Features biométricas
Informações sobre o contexto de aplicação</td><td>Saídas:
Modelagem do problema de aprendizado de máquina</td></tr>
</table>

Fonte – O autor.

4.3.3 Definir Variáveis de Resposta

Para Jain (1990), as variáveis de resposta são medidas de desempenho e representam os critérios usados para a avaliação de sistemas computacionais. Segundo o autor, para escolher as medidas corretas é preciso compreender os serviços fornecidos pelo sistema e o objetivo da avaliação. Uma solicitação de serviço de um sistema pode ser respondida com sucesso, erro ou pode ser recusada, sendo que as medidas de desempenho associadas a cada um desses casos são a velocidade, confiabilidade e disponibilidade, respectivamente. As medidas de velocidade estão relacionadas ao tempo de resposta do serviço (responsividade), taxa de resposta (produtividade) e recursos utilizados (utilização). Já as medidas de confiabilidade e disponibilidade buscam

classificar respectivamente os tipos de erros e falhas de serviços e determinar a probabilidade de ocorrência deles no sistema avaliado. Para muitas medidas, a análise deve considerar principalmente o valor médio obtido a partir de um conjunto adequado de medições, sem esquecer de levar em conta a variabilidade.

Nas avaliações de algoritmos de aprendizado de máquina, as medidas de confiabilidade (erro) representam um dos principais indicadores de desempenho (ALPAYDIN, 2014). Por isso, na subseção 2.2.2.3 deste trabalho são apresentadas várias medidas de erro (eficácia) para algoritmos de classificação two-class que representam os tipos de erros comumente considerados na avaliação desses tipos de classificadores. Além disso, em alguns casos, medir os recursos utilizados (eficiência) pelos algoritmos de aprendizado é tão importante quanto medir os seus erros (ALPAYDIN, 2014). Com base nisso, o Quadro 6 apresenta os detalhes da atividade Definir Variáveis de Resposta do processo proposto.

Quadro 6 – Detalhes da atividade Definir Variáveis de Resposta.

<table>
<tr><td colspan="2">Objetivo: Definir as medidas de desempenho que serão utilizadas como critério para a avaliação das features biométricas.</td></tr>
<tr><td colspan="2">Descrição: Nesta atividade devem ser definidas as medidas de desempenho que serão utilizadas como critério para a avaliação das features biométricas. Para isso, devem ser considerados principalmente o objetivo da avaliação e o tipo de algoritmo de aprendizado previsto na modelagem do problema. No entanto, é importante ter em mente que o objeto da avaliação previsto neste processo são as features biométricas e não o sistema biométrico como um todo. Assim, por exemplo, medidas relacionadas com a disponibilidade do sistema não são variáveis de resposta de interesse, já que a disponibilidade não é influenciada pelas features utilizadas.</td></tr>
<tr><td>Entradas:
Objetivo da avaliação
Modelagem do problema de aprendizado de máquina</td><td>Saídas:
Medidas de desempenho selecionadas</td></tr>
</table>

Fonte – O autor.

4.3.4 Definir Fatores e Níveis

Os fatores são variáveis que influenciam o desempenho do sistema e os níveis são valores que um mesmo fator pode assumir (JAIN, 1990). Por exemplo, o número de usuários conectados simultaneamente precisa ser considerado quando se pretende avaliar o tempo de resposta de um sistema, pois é esperado um aumento do tempo de resposta de um sistema com um aumento do número de usuários conectados simultaneamente nele. Nesse caso, a variável número de usuários conectados simultaneamente representa um fator e os valores 10, 100 e 1000 podem representar níveis representativos para esse fator no sistema avaliado.

Para Jain (1990), o desempenho de sistemas pode ser influenciado por vários fatores. No entanto, para ele, é importante identificar os fatores com impacto mais significativo no desempenho. Por outro lado, subestimar fatores ou níveis importantes pode tornar inútil o resultado da avaliação. Os conjuntos de dados utilizados também representam um exemplo de fator, pois impactam significativamente nos resultados da avaliação. Porém, a utilização de dados não representativos do uso real de um sistema torna inútil o resultado da sua avaliação.

Como já foi apresentado na subseção 2.2.2.1, segundo Alpaydin (2014), na avaliação de algoritmos de aprendizado existem fatores controláveis e incontroláveis. Nesta atividade, o foco é na definição dos fatores controláveis e seus níveis representativos para o problema. Com base nisso, o Quadro 7 apresenta os detalhes da atividade Definir Fatores e Níveis do processo proposto.

Quadro 7 – Detalhes da atividade Definir Fatores e Níveis.

<table>
<tr><td colspan="2">Objetivo: Definir os fatores e níveis que influenciam o desempenho das features biométricas e serão utilizados na avaliação.</td></tr>
<tr><td colspan="2">Descrição: Nesta atividade devem ser definidos os fatores e níveis que influenciam o desempenho das features biométricas e serão utilizados na avaliação. Para isso, devem ser consideradas a modelagem do problema de aprendizado de máquina e as medidas de desempenho selecionadas. A modelagem servirá para identificar os algoritmos que poderão ser utilizados na avaliação. As medidas representam os critérios de desempenho da avaliação e podem indicar outros fatores e níveis importantes.</td></tr>
<tr><td>Entradas:
Modelagem do problema de aprendizado de máquina
Medidas de desempenho selecionadas</td><td>Saídas:
Fatores e níveis selecionados</td></tr>
</table>

Fonte – O autor.

4.3.5 Executar Pré-testes

Nos pré-testes, algumas configurações aleatórias são selecionadas e testadas antes da execução de um grande experimento com muitos fatores e níveis para verificar se tudo está conforme o esperado (JAIN, 1990). No entanto, neste processo, a atividade Executar Pré-testes tem por objetivo aumentar a compreensão do avaliador sobre os fatores e níveis identificados para a avaliação. Segundo Jain (1990), um erro comum em projetos de avaliação é ignorar ou subestimar fatores ou níveis significativos para o desempenho do objeto avaliado. Por isso, essa atividade foi posicionada entre a definição dos fatores e níveis e a definição do design do experimento para evitar esses erros.

Em experimentos muito complexos essa etapa ganha maior relevância, já que nem

sempre o design fatorial será uma alternativa viável, conforme foi apresentado na subseção 2.2.2.2. Nesses casos, realizar os pré-testes com alguns cenários considerados importantes, e não apenas com cenários aleatórios, antes da definição do design do experimento, ajuda na compreensão dos fatores que influenciam significativamente o desempenho do sistema e como esses fatores interagem entre si. Com essas informações, busca-se estabelecer algumas premissas sobre os fatores e níveis, diminuir o número de experimentos necessários e, ao mesmo tempo, melhorar a qualidade dos resultados.

Com base no que foi discutido anteriormente, o Quadro 8 apresenta os detalhes da atividade Executar Pré-testes do processo proposto.

Quadro 8 – Detalhes da atividade Executar Pré-testes.

Objetivo: Melhorar a compreensão sobre como os fatores e níveis influenciam o desempenho das *features* biométricas.	
Descrição: Nesta atividade devem ser executados pré-testes para melhorar a intuição sobre como os fatores e níveis influenciam o desempenho das *features* biométricas. Para isso, devem ser considerados os fatores e níveis selecionados, principalmente os algoritmos e conjuntos de dados. Com isso, é possível priorizar os fatores e níveis, definir parâmetros para pré-processamento dos dados e otimizar o experimento planejado.	
Entradas: • Fatores e níveis selecionados	**Saídas:** • Fatores e níveis priorizados • Parâmetros para pré-processamento dos dados

Fonte – O autor.

4.3.6 Definir Design do Experimento

O design experimental refere-se ao número de experimentos de medição que devem ser realizados e os níveis utilizados em cada um deles (JAIN, 1990). Com os fatores e níveis selecionados, é preciso decidir entre inúmeras sequência de experimentos possíveis, priorizando as sequências que oferecem mais informações com menos esforço. A seleção adequada desses níveis pode gerar resultados com mais informações a partir do mesmo número de experimentos (JAIN, 1990). Por outro lado, a seleção inadequada pode resultar em um desperdício de recursos para o projeto, principalmente na execução dos experimentos e análise dos resultados (JAIN, 1990).

Como já foi apresentado na subseção 2.2.2.2, Alpaydin (2014) apresenta três estratégias para experimentos com aprendizado de máquina: melhor suposição; um fator de cada vez; e design fatorial. Segundo ele, é melhor usar o design fatorial, mas existem algumas exceções. Uma das exceções é quando se tem certeza de que alguns fatores não interagem entre si. Sendo

assim, os cenários envolvendo as combinações desses fatores não precisam ser considerados na avaliação. No entanto, na maioria das vezes, os fatores interagem e precisam ser avaliados (ALPAYDIN, 2014). Outra exceção é quando o número de fatores e níveis cresce muito e isso torna o problema não tratável computacionalmente.

Com base nisso, o Quadro 9 apresenta os detalhes da atividade Definir Design do Experimento do processo proposto.

Quadro 9 – Detalhes da atividade Definir Design do Experimento.

Objetivo: Definir a sequência de experimentos que devem ser executados, otimizando as informações geradas e os recursos utilizados.	
Descrição: Nesta atividade deve ser escolhida a estratégia de experimentação aplicável. A partir dela, deve ser definida a sequência de experimentos que serão executados. Para isso, devem ser considerados os fatores e níveis priorizados e os recursos disponíveis para a realização da avaliação. Assim, a sequência de experimentos definida deve gerar o máximo de informações com os recursos disponíveis. As informações são maximizadas priorizando a utilização dos fatores e níveis que mais influenciam o resultado.	
Entradas: • Fatores e níveis priorizados • Recursos disponíveis	**Saídas:** • Sequência de experimentos

Fonte – O autor.

4.4 Execução

Para Jain (1990), a execução dos experimentos deve ser conduzida por profissionais especializados no uso das técnicas e ferramentas de avaliação adequadas, observando os requisitos definidos no planejamento. Além disso, não podem existir preconceitos e preferências sobre o objeto da avaliação entre esses profissionais. Todos os experimentos devem ser realizados com o mesmo cuidado, pois uma regra de ouro para analistas de desempenho é ser imparcial (JAIN, 1990).

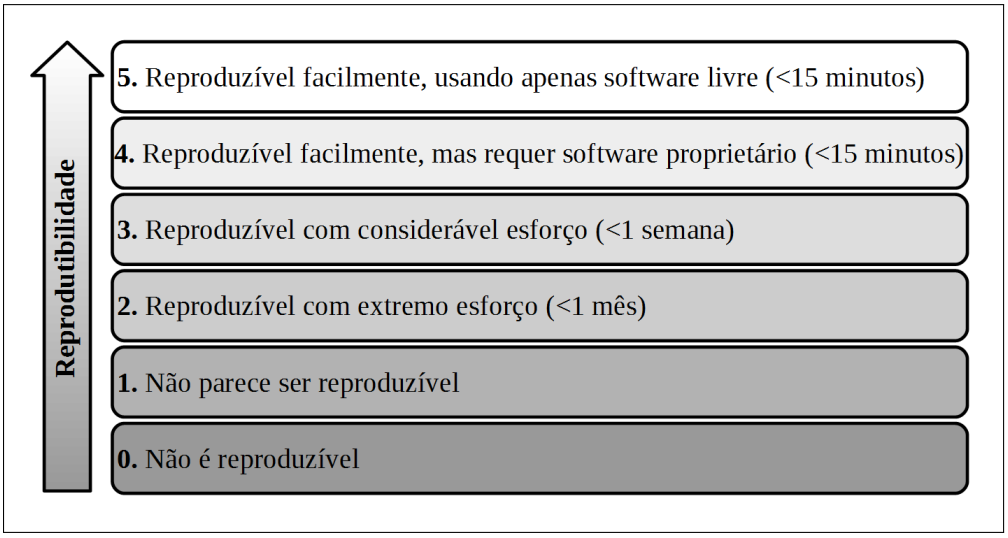
A reprodutibilidade do experimento é um aspecto importante que deve ser considerado ao máximo nessa atividade. Com base na definição de pesquisa reproduzível (VANDEWALLE *et al.*, 2009), um experimento é reproduzível neste trabalho se todas as informações relevantes para a execução do experimento são disponibilizadas, de modo que outra pessoa interessada pode reproduzir de maneira independente os resultados. Para isso, as ferramentas, os dados e os detalhes do experimento devem ser claramente definidos.

Na avaliação de *features* biométricas, a reprodutibilidade permite que outros pesquisadores possam validar de maneira independente os resultados de um experimento, compará-los

e reutilizar as *features* biométricas que oferecem melhor desempenho.

A Figura 8 apresenta uma escala baseada nos graus de reprodutibilidade da pesquisa científica (VANDEWALLE *et al.*, 2009). Ela busca apoiar as decisões dos profissionais especializados na execução no sentido de tornar os experimentos o mais reproduzível possível.

Figura 8 – Escala de reprodutibilidade do experimento.



Fonte – O autor, baseado em Vandewalle *et al.* (2009).

No caso do aprendizado de máquina, é fundamental usar ferramentas testadas, otimizadas e obtidas de fontes confiáveis (ALPAYDIN, 2014). Isso busca evitar problemas no código que podem afetar os resultados. Além disso, na execução de grandes experimentos é recomendado sempre salvar resultados intermediários e sementes do gerador de números aleatórios (ALPAYDIN, 2014). Assim, o experimento pode ser repetido parcialmente ou por completo, caso seja necessário.

Ainda no caso de experimentos com aprendizado de máquina, é importante utilizar técnicas que garantam os princípios estatísticos da repetição (*replication*), aleatoriedade (*randomization*) e blocagem (*blocking*) (ALPAYDIN, 2014). Como foi apresentado nas subseções 2.2.2.4 e 2.2.2.5, os experimentos precisam ser repetidos para que o efeito dos fatores incontroláveis seja anulado (repetição). Além disso, os dados precisam ser aleatoriamente reordenados para que os resultados obtidos nas diferentes repetições sejam independentes (aleatoriedade). Por fim, para comparar dois ou mais algoritmos de aprendizado os dados utilizados devem ser exatamente os mesmos para evitar a mudança de fatores indesejados (blocagem). Mais informações sobre as várias técnicas que podem ser usadas para garantir esses princípios estatísticos em experimentos de aprendizado de máquina e as situações mais adequadas para a utilização de cada uma delas

podem ser encontradas em Alpaydin (2014) e também em alguns trabalhos específicos sobre o assunto (BOUCKAERT; FRANK, 2004)(DEMŠAR, 2006)(DIETTERICH, 1998).

Com base nisso, o Quadro 10 apresenta os detalhes da atividade Executar Experimento do processo proposto.

Quadro 10 – Detalhes da atividade Executar Experimento.

Objetivo: Selecionar as técnicas e ferramentas que serão utilizadas, realizar o pré-processamento dos dados e executar o experimento.	
Descrição: Nesta atividade devem ser selecionadas as técnicas e ferramentas que serão utilizadas para a execução do experimento. Para isso, deve ser considerado o planejamento da avaliação. Além disso, antes da execução do experimento deve ser realizado o pré-processamento dos conjuntos de dados selecionados de acordo com os parâmetros definidos na atividade pré-testes. Por fim, devem ser executados os experimentos de acordo com a sequência definida na atividade design do experimento.	
Entradas: • Objetivo da avaliação • Modelagem do problema de aprendizado de máquina • Medidas de desempenho selecionadas • Fatores e níveis priorizados • Parâmetros para pré-processamento dos dados • Sequência de experimentos	**Saídas:** • Resultados do experimento

Fonte – O autor.

4.5 Análise

Para Jain (1990), um projeto de avaliação de desempenho deve contar também com profissionais especializados em análise de dados. Segundo ele, um dos problemas comuns em projetos de avaliação é que eles são freqüentemente executados por analistas de desempenho que são bons em técnicas de medição, mas não têm experiência em análise de dados. Análises realizadas por profissionais não especializados nessa área apresentam frequentemente três tipos de problemas:

- Nenhuma análise. São apresentados apenas resultados brutos e gráficos incompreensíveis, sem análise ou explicações detalhadas de como eles podem ser utilizados para extrair informações úteis.
- Análise errônea. A análise dos resultados é feita utilizando técnicas inadequadas. Por exemplo: o cálculo da média de porcentagens, desconsiderando as proporções originais que deram origem a elas; e a análise apenas do desempenho médio, desconsiderando a variabilidade.
- Nenhuma análise de sensibilidade. Os resultados podem ser sensíveis ao conjunto

de dados e parâmetros do sistema. Sem uma análise dessa sensibilidade não é possível garantir que as conclusões serão as mesmas em cenários ligeiramente diferentes.

Na análise dos resultados, ao sumarizar e comparar o desempenho de sistemas usando medidas de tendência central (*e.g.* média, moda, mediana, entre outras) é importante considerar também a necessidade do uso de medidas de variabilidade (*e.g.* variância, desvio padrão, entre outras) (JAIN, 1990). Por exemplo, o uso da média aritmética sem considerar a variância pode levar a conclusões incorretas. As principais técnicas estatísticas utilizadas para comparar o desempenho de dois sistemas são o intervalo de confiança e o teste de hipótese (ALPAYDIN, 2014)(JAIN, 1990). Essas técnicas permitem que as conclusões obtidas a partir da análise dos resultados do experimento sejam objetivas e tenham nível de erro conhecido e controlado (ALPAYDIN, 2014). Assim, as questões que uma avaliação precisa responder devem ser convertidas para uma estrutura de teste de hipótese. Em seguida deve ser verificado se os resultados do experimento apóiam essas hipóteses. Mais informações sobre as várias técnicas de comparação de desempenho usadas em experimentos de aprendizado de máquina e as situações mais adequadas para a utilização de cada uma delas podem ser encontradas em Alpaydin (2014).

Com base no que foi discutido anteriormente, o Quadro 11 apresenta os detalhes da atividade Analisar Resultados do processo proposto.

Quadro 11 – Detalhes da atividade Analisar Resultados.

Objetivo: Analisar estatisticamente os resultados do experimento e comparar o desempenho das *features* nos cenários avaliados.	
Descrição: Nesta atividade, os resultados do experimento devem ser analisados estatisticamente. Para isso, os resultados brutos obtidos para uma medida de desempenho devem ser sumarizados, considerando um mesmo cenário avaliado (*features* biométricas e os demais fatores e níveis selecionados). Depois disso, essas medidas sumarizadas devem ser comparadas, considerando dessa vez os diferentes cenários avaliados. Em seguida, devem ser identificados os cenários com diferenças de desempenho significativas entre eles. Por fim, deve ser analisada a influência dos fatores e níveis envolvidos nesses cenários, focando principalmente nas *features* biométricas e nos cenários em que elas oferecem melhor desempenho.	
Entradas: • Resultados do experimento	**Saídas:** • Resultados da análise

Fonte – O autor.

4.6 Finalização

Para Jain (1990), projetos de avaliação de desempenho devem contar também com profissionais com habilidades técnicas (*e.g.* analisar dados) e sociais (*e.g.* escrever, falar, entre

outras) para a documentação dos resultados. As habilidades técnicas ajudam a interpretar os resultados da análise corretamente. Já as habilidades sociais ajudam a transmitir os resultados para o público interessado de maneira adequada. Segundo Jain (1990), analistas inexperientes frequentemente:

- Têm dificuldade de compreender a importância das habilidades sociais;
- Não dedicam esforço e tempo suficiente na apresentação dos resultados;
- Apresentam os resultados sem considerar as características do público interessado, utilizando palavras, figuras, gráficos e jargões inadequados;
- Apresentam os resultados de uma mesma forma buscando atender vários públicos com características distintas, tornando a apresentação sem sentido para todas essas audiências; e
- Estão mais interessados em apresentar os detalhes técnicos do projeto e as inovações utilizadas do que os resultados finais.

É importante que os resultados sejam apresentados de uma maneira que seja facilmente compreensível. Isso geralmente requer o uso de gráficos adequados (ALPAYDIN, 2014)(JAIN, 1990). No entanto, o objetivo dos gráficos é apenas complementar a informação e facilitar a compreensão dos resultados. Eles não retiram a exigência de existirem explicações detalhadas sobre como os resultados podem ser interpretados. Mais informações sobre os vários tipos de gráficos usados na avaliação de desempenho, as situações mais adequadas para a utilização de cada uma delas e um checklist com boas práticas sobre esse assunto podem ser encontradas em Jain (1990).

Além disso, uma vez que todos os dados são coletados e analisados, pode-se construir conclusões objetivas. Sobre isso é importante lembrar que os testes estatísticos nunca nos dizem se a hipótese é correta ou falsa. Eles apenas indicam o quanto a amostra parece concordar com a hipótese (ALPAYDIN, 2014). Existe sempre o risco de não termos um resultado conclusivo ou de que nossas conclusões estejam erradas, especialmente se os dados forem pequenos ou tiverem muito ruído. Assim, uma conclusão que precisa ser sempre considerada é a necessidade de novas experiências. Isso é normal e acontece com frequência, pois a maioria dos estudos estatísticos são iterativos (ALPAYDIN, 2014)(JAIN, 1990).

Com base nisso, a Quadro 12 apresenta os detalhes da atividade Apresentar Resultados do processo proposto.

Quadro 12 – Detalhes da atividade Apresentar Resultados.

Objetivo: Preparar os resultados da análise para a apresentação ao público interessado.	
Descrição: Essa atividade deve ter início pela identificação do público interessado no resultado da avaliação e as características dessa audiência. Com base nisso, os resultados da análise precisam ser interpretados e transformados em gráficos, tabelas, descrições detalhadas das evidências encontradas e conclusões objetivas. Por fim, deve ser gerado o relatório da avaliação específico para o público identificado. Caso sejam identificadas audiências com características diferentes, devem ser gerados relatórios específicos para cada uma delas.	
Entradas: • Resultados da análise	**Saídas:** • Relatórios da avaliação

Fonte – O autor.

4.7 Considerações Finais

Neste capítulo foi apresentado o processo de avaliação de *features* biométricas elaborado com base em recomendações de boas práticas para avaliação de desempenho de sistemas computacionais (JAIN, 1990) e experimentos de aprendizado de máquina (ALPAYDIN, 2014).

Esse processo foi modelado utilizando a notação BPMN (CBOK, 2013) e é composto por nove atividades agrupadas em quatro especialidades. Essas atividades foram inspiradas nas recomendações de boas práticas utilizadas e nas necessidades específicas percebidas na avaliação de *features* biométricas para autenticação transparente e contínua em dispositivos móveis, usando aprendizado de máquina. Já as especialidades foram inspiradas nos erros frequentes identificados e nas boas práticas recomendadas para avaliação de desempenho de sistemas computacionais (JAIN, 1990). Além disso, para cada atividades desse processo foram definidos também o objetivo, a descrição, as entradas e as saídas.

Para testar a aplicabilidade do processo proposto, um estudo de caso para a avaliação de *features* de localização é conduzido e os detalhes são apresentados no Capítulo 5.

5 AVALIAÇÃO DE *FEATURES* DE LOCALIZAÇÃO

Neste capítulo são descritos os detalhes da avaliação de *features* de localização realizada durante esta pesquisa. Essa avaliação foi conduzida fazendo uso do processo proposto neste trabalho e das *features* de localização utilizadas nos trabalhos relacionados identificados na revisão da literatura apresentada na seção 3.2.

As seções deste capítulo estão estruturadas da seguinte forma: a seção 5.1 apresenta uma breve introdução com a justificativa das principais escolhas tomadas para a realização dos experimentos; na sequência, as seções 5.2 a 5.5 descrevem os detalhes da avaliação relacionados a cada um dos conjuntos de atividades previstos no processo proposto; e, por fim, a seção 5.6 apresenta as considerações finais deste capítulo.

5.1 Introdução

O objetivo principal dessa avaliação foi produzir evidência sobre a viabilidade do processo proposto. Além disso, ela também produziu resultados que permitem identificar as *features* de localização *outdoor* e os algoritmos de aprendizado que oferecem melhor desempenho para a construção de mecanismos de autenticação transparente e contínua. Esses resultados são úteis em pesquisas que buscam usar *features* de localização *outdoor* combinadas com *features* extraídas de outros traços biométricos para compor novos sistemas multibiométricos de autenticação. Além disso, elas são úteis também em pesquisas que buscam propor novas *features* de localização *outdoor* e precisam de referências de desempenho para comparação.

A escolha das *features* de localização *outdoor* para a avaliação do processo proposto levou em consideração o número significativo de propostas encontradas na literatura que utilizam a localização *outdoor* como *features* biométricas para autenticação transparente e contínua. Além disso, considerou também as deficiências encontradas nas avaliações dessas propostas e a quantidade representativa de dados disponíveis publicamente de localização *outdoor* obtida de usuários reais em situações normais de utilização de dispositivos móveis. Ademais, as técnicas de localização *outdoor* podem ser usadas na grande maioria dos dispositivos móveis atuais, permitindo a utilização em larga escala e com baixo custo.

5.2 Planejamento

Nesta seção são descritos os detalhes do conjunto de atividades de planejamento previstas no processo proposto, apresentado no Capítulo 4.

5.2.1 Definir Objetivo

O objetivo desta avaliação é comparar a eficácia e a eficiência das *features* de localização utilizadas nas propostas de mecanismos de autenticação transparente e contínua para dispositivos móveis disponíveis na literatura, produzindo evidências que permitam determinar as suas vantagens e desvantagens.

Nesta avaliação, conforme apresentado na seção 1.1, a eficácia diz respeito à produção do resultado esperado com ausência de erros (MICHAELIS, 2015). Essa definição está relacionada com as medidas de confiabilidade na avaliação de desempenho (JAIN, 1990). Essas medidas buscam classificar os erros possíveis em um sistema computacional e determinar as suas respectivas probabilidades de ocorrência. No aprendizado de máquina, as medidas de desempenho mais utilizadas para avaliar algoritmos de classificação e regressão envolvem o erro (ALPAYDIN, 2014). Já a eficiência diz respeito aos recursos computacionais utilizados, seja na avaliação de sistemas computacionais (JAIN, 1990), como também no projeto de algoritmos (CORMEN, 2009) e no aprendizado de máquina (ALPAYDIN, 2014). Assim, é considerado mais eficiente o sistema ou algoritmo que utiliza menos recursos computacionais para a produção do resultado esperado.

5.2.2 Modelar Problema

Na avaliação das *features* de localização, o problema da autenticação de usuários foi modelado como um problema de aprendizagem supervisionada, usando classificadores de duas classes *two-class* (ALPAYDIN, 2014). Nessa modelagem, as amostras são compostas por um conjunto de *features* de localização e uma classe que determina se a amostra representa o comportamento do usuário autorizado (classe positiva) ou não (classe negativa). Essas amostras são usadas para treinar um classificador que depois deve ser capaz de fazer predições a partir de novas amostras, indicando se elas combinam com o padrão aprendido para a classe positiva ou negativa. Essas novas amostras devem ser diferentes das usadas no treinamento (inéditas) e ainda não classificadas. Essa modelagem foi escolhida com base nos trabalhos relacionados

apresentados na seção 3.2 deste trabalho. Além disso, essa modelagem também é compatível com os exemplos de contextos de aplicações de sistemas biométricos apresentados em Jain *et al.* (2007) e Alpaydin (2014).

Esse problema de autenticação também pode ser modelado de outra forma, usando classificadores one-class (ALPAYDIN, 2014). Nessa outra modelagem, as amostras disponíveis para o treinamento do classificador são apenas positivas. Baseado nisso, o classificador deve ser capaz de indicar se uma amostra inédita e ainda não classificada é compatível com o padrão aprendido para a classe positiva ou não (negativa). Essa modelagem não foi escolhida porque ela restringe muito o número de algoritmos disponíveis no ambiente WEKA (HALL *et al.*, 2009) capazes de resolver esse problema de aprendizado de máquina. Além disso, como já foi esclarecido nesta subseção, os exemplos didáticos encontrados na literatura e os trabalhos relacionados identificados usam a mesma modelagem utilizada neste trabalho. Mesmo assim, é importante destacar que alguns testes foram feitos durante esta pesquisa usando essa modelagem que utiliza apenas amostras positivas, o ambiente WEKA (HALL *et al.*, 2009) e o algoritmo LibSVM (CHANG; LIN, 2011) no modo *one-class* SVM. No entanto, os resultados obtidos nesses testes foram muito inferiores aos resultados apresentados nos trabalhos relacionados e também aos resultados obtidos nesta avaliação usando classificadores *two-class*, com resultados próximos aos obtidos usando estratégias de adivinhação.

5.2.3 Definir Variáveis de Resposta

Nessa avaliação foram utilizadas nove medidas de eficácia e três de eficiência. Para medir a eficácia foram utilizadas as medidas de erro propostas por Alpaydin (2014), apresentadas na subseção 2.2.2.3 deste trabalho (Acurácia, TPR, FPR, AUC, *Precision*, *Recall*, *F-Measure*, *Sensibility*, *Specificity*). Elas foram utilizadas porque são as mais frequentemente consideradas em problemas de classificação, especialmente *two-class* (ALPAYDIN, 2014). Essas medidas foram selecionadas também porque são equivalentes às medidas comumente utilizadas para a avaliação de sistemas biométricos, conforme apresentado na subseção 2.1.2.4 deste trabalho. Já para medir a eficiência foram utilizadas medidas relacionadas com a utilização de recursos computacionais, mais especificamente a memória utilizada para armazenamento do modelo classificador gerado e o tempo de CPU utilizado para o treinamento e o teste do modelo. Outras medidas de utilização de recursos computacionais também são importantes na avaliação de aplicações para dispositivos móveis (*e.g.* Memória RAM, bateria, entre outras) (MAIA *et al.*,

2013)(KHAN *et al.*, 2014)(BUTHPITIYA *et al.*, 2014). No entanto, neste trabalho foram usadas apenas as medidas suportadas pelo ambiente WEKA (HALL *et al.*, 2009). Mais informações sobre as técnicas utilizadas na coleta e análise das medidas no ambiente WEKA (HALL *et al.*, 2009) são apresentados nas Seções 5.2 e 5.3, que tratam respectivamente das atividades Executar Experimento e Analisar Resultados do processo proposto neste trabalho.

5.2.4 Definir Fatores e Níveis

Os conjuntos de *features* de localização representam o principal fator para este trabalho. Eles foram elaborados a partir dos trabalhos relacionados apresentados na seção 3.2 e são definidos a seguir:

- Features Set #1: (local1)
- Features Set #2: (local1, time1)
- Features Set #3: (local1, local2, local3, local4, local5, local6)
- Features Set #4: (local1, local2, local3, local4, local5, local6, time6)

Os elementos que compõem esses conjuntos são do tipo "local" ou "time". Nesta avaliação, os elementos do tipo "local" representam um subconjunto de *features* de localização, podendo ser uma coordenada geográfica (*i.e.* latitude e longitude) obtida por meio de GPS ou um identificador de célula (*i.e.* cell_area e cell_id) obtido por meio do Sistema de Telefonia Celular. Já os elementos do tipo "time" representam uma *feature* com o tempo em segundos desde o início do dia até a chegada na respectiva localização, variando de 0 à 86.399. Os dois tipos de elementos são seguidos de um índice que representa a ordem entre os elementos do mesmo tipo e a ligação entres os elementos de tipos diferentes. Assim, local1 e local2 representam duas localizações diferentes, sendo que local2 foi visitado imediatamente depois de local1. Já time6 representa o instante de tempo do dia que local6 foi visitado.

Como pode ser visto, o conjunto de *features* denominado Features Set #1 representa uma amostra com apenas uma localização visitada e Features Set #2 inclui ainda o período do dia que ocorreu essa visita. Já o conjunto Features Set #3 representa uma amostra com o rastro das últimas 6 localizações visitadas. Da mesma forma, Features Set #4 inclui ainda o período do dia que ocorreu a visita à última localização. O tamanho da janela deslizante da sequência de localizações usada em Features Set #3 e Features Set #4 foi fixado em 6, seguindo o resultado observado por Shi *et al.* (2011). Esses conjuntos de *features* foram definidos a partir das descrições fornecidas nos trabalhos relacionados. No entanto, essas descrições não são claras

e detalhadas. Por isso, não é possível ter uma compreensão completa das *features* de localização realmente utilizadas nesses trabalhos. Assim, esses conjuntos de *features* definidos para esta avaliação representam uma interpretação das descrições fornecidas pelos autores dos trabalhos relacionados e, ao mesmo tempo, uma proposta de implementação.

Os algoritmos de aprendizado selecionados também são fatores importantes para este trabalho. Nesta avaliação foram utilizados os algoritmos classificadores de aprendizado supervisionado de máquina utilizados com *features* de localização nos trabalhos relacionados apresentados na seção 3.2 e suportados no ambiente WEKA. Além disso, foram considerados também os algoritmos classificadores de aprendizado supervisionado de máquina utilizados com outros tipos de *features* nos trabalhos relacionados descritos por Patel *et al.* (2016) e Khan (2016). Assim, os algoritmos selecionados foram os seguintes:

- DT: Decision Tree C4.5 (QUINLAN, 2014)
- SVM: Support Vector Machine C-SVC (CHANG; LIN, 2011)
- NB: Naive Bayes (JOHN; LANGLEY, 1995)
- 0R: ZeroR[1]

O classificador 0R foi utilizado para estabelecer uma linha de base comparativa. Ele calcula a classe nominal mais frequente das amostras (moda) utilizadas na fase de treinamento para responder às predições na fase de teste. Isso permite estabelecer um desempenho mínimo esperado para os demais classificadores considerados.

Todos os algoritmos selecionados foram utilizados na avaliação com a configuração padrão no ambiente WEKA. Isso foi necessário, pois, caso contrário, envolveria um outro problema complexo, denominado CASH (THORNTON *et al.*, 2013), que envolve a seleção de algoritmos combinado com a otimização de configurações. Esse problema limitaria o design do experimento e os fatores e níveis considerados nesta avaliação, pois elevaria o número de experimentos necessários ao ponto de tornar a avaliação inviável em termos computacionais. Assim, a partir dos resultados gerados por este trabalho, novas pesquisas podem focar no problema da otimização das configurações dos algoritmos avaliados.

O SVM foi o único algoritmo que precisou receber ajuste na configuração padrão. Para esse algoritmo foi necessário ativar a normalização das *features* de entrada (opção -Z). Para Hsu *et al.* (2003), a padronização da escala (*i.e. Scaling*) dos dados antes de aplicar o SVM é muito importante, pois evita que as *features* com maiores intervalos numéricos dominem aqueles

[1] http://weka.sourceforge.net/doc.dev/weka/classifiers/rules/ZeroR.html

com intervalos numéricos menores e evita também dificuldades numéricas durante os cálculos. Além disso, foi desativado o uso de heurísticas (opção -H), pois o ganho de desempenho em termos de eficiência é mínimo, ao custo do aumento do número de erros (HSU *et al.*, 2003). No entanto, esses ajustes foram pontuais e estáticos, buscando apenas avaliar todos os algoritmos em condições de igualdade. Além disso, esses ajustes não representam uma tentativa de atacar o problema da otimização de configurações, pois isso exigiria testes com várias combinações dos diferentes hiperparâmetros de configuração específicos desse algoritmo.

Os conjuntos de dados de teste selecionados também são fatores importantes para este trabalho. Nesta avaliação foram utilizados dados disponibilizados publicamente e referenciados nos trabalhos relacionados identificados e descritos em Patel *et al.* (2016) e Khan (2016). Além disso, foram utilizados também dados não disponibilizados publicamente, coletados de usuários do aplicativo móvel Eai[2] e de voluntários desta pesquisa. Esses dados são apresentados no Quadro 13 com suas principais características.

Quadro 13 – Conjuntos de dados selecionados para os experimentos e as suas características

Cj. de Dados	**Dados Públicos**	**Região Predominante**	**Nº de Usuários**	**Tipo**
Geolife (ZHENG *et al.*, 2010)	Sim	Pequim, China	182	GPS
MIT Reality Mining (EAGLE; PENTLAND, 2006)	Sim	Massachusetts, Estados Unidos	94	Celular
Eai (DARIN *et al.*, 2016)	Não	Fortaleza, Brasil	411	GPS
Voluntários	Não	Fortaleza, Brasil	2	GPS

Fonte – O autor.

Os conjuntos de dados Geolife, Eai e Voluntários são formados por dados coletados via GPS (*i.e.* latitude, longitude e timestamp). Já o MIT é formado por dados coletados via Sistema de Telefonia Celular (*i.e.* cell_area, cell_id, timestamp). Os conjuntos de dados Eai e Voluntários foram obtidos diretamente dos dispositivos dos usuários participantes. No caso do Eai, os dados foram coletados dos usuários do aplicativo social móvel de mesmo nome. Já no caso dos Voluntários, os dados foram obtidos do histórico de localização do Google, gravado pelo sistema Android dos dispositivos. Assim, todos os dados utilizados neste trabalho são dados reais, coletados de usuários de dispositivos móveis.

O conjunto de dados Voluntários representa para este trabalho uma linha de base comparativa, pois os dois usuários possuem um grau considerável e conhecido de mobilidade

[2] https://play.google.com/store/apps/details?id=br.ufc.appeai&hl=pt_BR

urbana, regularidade dos hábitos de localização e semelhança entre os seus hábitos. Nesta avaliação, a mobilidade do usuário de forma regular dentro de uma região geográfica é fundamental para a produção de dados e padrões suficientes para o treinamento e teste dos algoritmos de aprendizado de máquina. Usuários estáticos ou com mobilidade aleatória não se beneficiam da autenticação transparente e contínua baseada em localização. Assim, é esperado que essa característica conhecida dos Voluntários favoreça os resultados obtidos usando esses dados. Por outro lado, usuários com hábitos de localização semelhantes produzem padrões similares de mobilidade, dificultando a diferenciação desses usuários. Já nesse caso, é esperado que essa outra característica conhecida dos Voluntários desfavoreça os resultados obtidos nos experimentos usando esses dados. Por fim, baseado nos resultados obtidos com esse conjunto de dados e considerando todas essas características conhecidas, busca-se compreender melhor as características dos usuários contidos nos outros conjuntos de dados utilizados nesta avaliação.

Mais informações sobre os conjuntos de dados Geolife (ZHENG *et al.*, 2010), MIT (EAGLE; PENTLAND, 2006) e Eai (DARIN *et al.*, 2016) são disponibilizados pelos seus respectivos autores e podem ser obtidas nas referências indicadas neste trabalho.

5.2.5 Executar Pré-testes

Nesta atividade foram feitos testes de pré-processamento dos dados para a extração das *features* de localização definidas na subseção 5.2.4. Além disso, foram conduzidos alguns experimentos de teste para compreender o funcionamento de algumas ferramentas e classificadores.

Primeiramente, percebeu-se que as amostras produzidas a partir dos conjuntos de dados obtidos via GPS apresentavam muitas localizações repetidas em sequência. Isso ocorria de acordo com a frequência de coleta específica de cada usuário e conjunto de dados. Nos primeiros testes de pré-processamento essas repetições foram mantidas também nas amostras de *features* usadas nos experimentos de teste para o treinamento e teste de algoritmos classificadores. No entanto, foi possível perceber que essas repetições tinham um grande impacto negativo na eficiência dos algoritmos de aprendizado, sem melhorar a sua eficácia. Por isso, foram buscadas estratégias de sumarização desses dados baseados em GPS. Os conjuntos de dados baseados em Celular não apresentam repetição e, por isso, não precisaram ser sumarizados.

Para essa sumarização, foi utilizada a discretização do espaço como em Tang *et al.* (2010) e Rocha *et al.* (2016). De forma semelhante, foi utilizada também a discretização do

tempo. Para isso, foram aplicadas escalas nos testes de pré-processamento, quando os dados brutos coletados (*i.e.* latitude, longitude e timestamp) são transformados em amostras com base nos conjuntos de *features* definidos na subseção 5.2.4. Durante esse pré-processamento, duas amostras consecutivas são registradas quando elas diferem por localização ou tempo, considerando a escala adotada. Isso quer dizer que a segunda amostra é descartada se dois registros de localização consecutivos representam localizações dentro de uma mesma região do espaço e um mesmo período de tempo, simultaneamente. Nos casos dos conjuntos de *features* que não incluem o tempo, as amostras são descartadas considerando apenas a localização.

A escala de espaço foi implementada por meio do arredondamento das coordenadas geográficas (latitude e longitude) em graus. Ela foi fixada com precisão de três casa decimais ($0{,}001^{o}$). Isso representa regiões do espaço com aproximadamente 100x100 metros nas referidas regiões onde os respectivos dados foram predominantemente coletados, conforme descrito no Quadro 13. Já a escala de tempo foi fixada em 1 minuto. Esses valores foram escolhidos durante os pré-testes, pois demonstraram equilibrar a granularidade e a perda de informação. Essas escalas também foram usadas para produzir as sequências de localização para as amostras do Features Set #3 e Features Set #4. Essas sequências são formados por localizações diferentes das suas vizinhas imediatas, considerando a mesma escala de espaço descrita.

Outra questão importante identificada nos testes foi que esses conjuntos com dados obtidos via GPS apresentam regiões predominantes onde as localizações coletadas dos usuários estão concentradas. No caso do Geolife, a região predominante é Pequim, na China (latitude: 39.440 a 41.060; longitude: 115.410 a 117.400). Já no caso dos conjuntos de dados Eai e Voluntários, a região predominante é Fortaleza, no Brasil (latitude: -3.888 a -3.691; longitude: -38.637 a -38.400). Assim, esta avaliação só deve considerar o desempenho na diferenciação dos padrões de mobilidade dos usuários com localizações dentro de uma mesma região, pois isso representa um modelo de ameaça realista para o problema de autenticação analisado. Além disso, serão desconsideradas em cada um desses conjuntos de dados as localizações fora dos limites definidos para as suas respectivas regiões predominantes identificadas.

Por fim, mais uma questão importante identificada nos testes foi a quantidade de dados necessária para o treinamento do algoritmo classificador. Nesse sentido, foram analisadas a quantidade necessária de amostras para o treinamento e de tempo para a coleta dos dados de localização. O classificador precisa de uma quantidade mínima de amostras de treinamento para identificar os padrões do usuário e fazer predições com uma taxa de erro aceitável. O treinamento

de classificadores com uma quantidade insuficiente de amostras é uma das causas de *underfitting* (ALPAYDIN, 2014). Essas amostras são geradas a partir do pré-processamento dos dados de localização coletados dentro de um intervalo de tempo compreendido entre a primeira e a última coletada.

Baseado nos pré-testes, o intervalo de tempo para obtenção dos dados de localização foi fixado em um período de 7 dias. Os pré-testes demonstraram que esse período permite o aprendizado de padrões diários e semanais dos usuários, que são importantes para os conjuntos de *features* que incluem o tempo. Além disso, na prática, esse período de treinamento é considerável para um potencial usuário de autenticação transparente e contínua. Já a quantidade mínima de amostras foi fixada em 100. Os pré-teste demonstraram que um número de amostras menor que esse limite provocava *underfitting* e aumentava muito a taxa de erro. Foram analisados alguns exemplos de usuários cujos dados coletados no período de 7 dias produziram menos de 100 amostras em todos os conjuntos de *features*. A partir disso, foi possível perceber que esses usuários apresentaram mobilidade urbana atípica (comportamento estático ou praticamente estático) no período analisado, e por isso são incompatíveis com a autenticação transparente e contínua baseada em localização.

5.2.6 Definir Design do Experimento

Para alcançar o objetivo deste trabalho foi escolhido o design fatorial para o experimento, gerando 64 cenários de avaliação (4 conjuntos de dados x 4 conjuntos de *features* x 4 algoritmos de aprendizado). Cada um desses cenários é composto por vários experimentos, sendo um para cada usuário considerado no respectivo conjunto de dados, de acordo com os parâmetros definidos na subseção 5.2.5. A Tabela 1 apresenta o número de experimentos por cenário e também o total de experimentos planejados para esta avaliação.

5.3 Execução

O ambiente WEKA foi escolhido para execução dos experimentos nesta pesquisa. As principais características que justificaram essa escolha são apresentadas na subseção 2.2.2.6 deste trabalho. Além disso, o WEKA demonstrou ao longo desta pesquisa ser mais fácil de utilizar do que a outra alternativa disponível analisada, a biblioteca python para aprendizado de máquina denominada scikit-learn[3]. Assim como as outras características citadas, a facilidade é

[3] http://scikit-learn.org/stable/

Tabela 1 – Experimentos planejados

Cj. de Dados	**Cj. de *Features***	**Nº de Experimentos**
Geolife	Features Set #1	448
	Features Set #2	464
	Features Set #3	448
	Features Set #4	464
MIT	Features Set #1	280
	Features Set #2	280
	Features Set #3	280
	Features Set #4	280
Eai	Features Set #1	264
	Features Set #2	564
	Features Set #3	264
	Features Set #4	564
Voluntários	Features Set #1	8
	Features Set #2	8
	Features Set #3	8
	Features Set #4	8
Total de experimentos		4632

Fonte – O autor.

uma questão fundamental para a reprodutibilidade dos experimentos, conforme foi discutido na seção 4.4.

A execução do experimento utilizou o ambiente WEKA 3.7.13, com pacote LibSVM 1.0.8 instalado por meio do gerenciador de pacotes incluído nesse ambiente. O equipamento utilizado foi um PC Lenovo ThinkCenter M91p com processador Intel® Core™ i5-2400, 8GB de memória RAM, sistema operacional Microsoft Windows 7 Professional 64bits SP1 e plataforma Java Oracle 1.8.0.

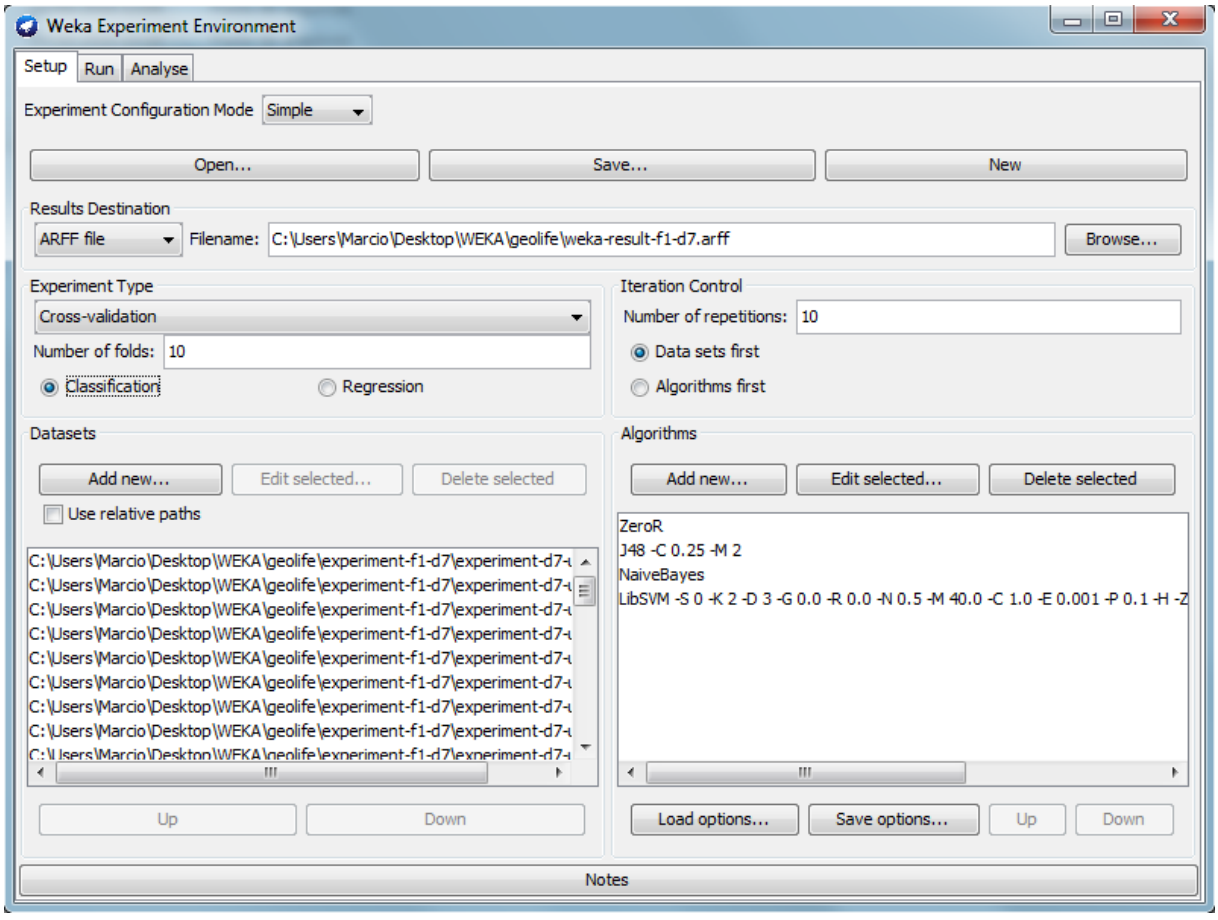
Na avaliação do processo proposto utilizando esta plataforma computacional foram executados no total 463.200 experimentos em 44 horas, 40 minutos e 7 segundos, considerando todos os cenários (4 conjuntos de dados x 4 conjuntos de *features* x 4 algoritmos de aprendizado x 10x10-*Fold* CV).

Para a execução dos experimentos planejados foi necessário nessa atividade implementar as rotinas de pré-processamento para geração dos arquivos no formato ARFF suportados pelo ambiente WEKA. Esse pré-processamento foi baseado na modelagem de problema de aprendizado de máquina descrita na subseção 5.2.2, nos conjuntos de *features* e de dados apresentados na subseção 5.2.4, nos parâmetros para pré-processamento descritos na subseção 5.2.5 e no design do experimento apresentado na subseção 5.2.6. Os arquivos ARFF são formados por amostras com um conjunto de *features* e uma classe. Para a geração desses arquivos foram

implementadas rotinas de pré-processamento em Python que recebem os conjuntos de dados selecionados e os transformam em amostras balanceadas das classes positivo e negativo, considerando os cenários definidos, os usuários do conjunto de dados e os conjuntos de *features* definidos. As amostras de classe positivas são do usuário alvo do experimento e as negativas são dos demais usuários daquele mesmo conjunto de dados. Os arquivos ARFF gerados foram usados para o treinamento e teste dos algoritmos de aprendizado selecionados.

A Figura 9 apresenta a interface para configuração de experimentos do WEKA. Nessa figura, a interface está exibindo as configurações para execução dos experimentos envolvendo o conjunto de dados Geolife e o conjunto Features Set #1, conforme apresentado na Tabela 1 da subseção 5.2.6. Nessa configuração é possível definir os algoritmos (*Algorithms*), os arquivos ARFF (*Datasets*), o número de subconjuntos (*Number of folds*) e o número de repetições (*Number of repetitions*) para a execução dos experimentos.

Figura 9 – Interface de configuração do experimento do WEKA



Fonte – O autor.

É importante destacar que a configuração para o número de repetições e o número

de conjuntos da validação cruzada (*Cross Validation* - CV) representa uma questão investigada em alguns trabalhos na área de aprendizado de máquina, mas que permanece em aberto. Isso ocorre porque alguns desses trabalhos buscam apenas minimizar a probabilidade de erros na detecção de diferenças significativas entre os resultados obtidos e para isso indicam o uso de uma configuração 5x2-*Fold* CV (DIETTERICH, 1998)(ALPAYDM, 1999). Já outros buscam, além disso, aumentar a reprodutibilidade dos resultados dos experimentos e para isso indicam o uso de uma configuração 10x10-*Fold* CV (BOUCKAERT; FRANK, 2004). Assim, nesta avaliação foi utilizada a configuração 10x10-*Fold* CV, pois ela foi proposta pelos próprios autores do WEKA (HALL *et al.*, 2009) e é sugerida por padrão nesse ambiente. Além disso, segundo os resultados obtidos em (BOUCKAERT; FRANK, 2004), essa configuração aumenta a reprodutibilidade do experimento e isso está alinhado com os objetivos deste trabalho.

Por fim, em relação à plataforma computacional para a execução destes experimentos, vale ressaltar que, embora o presente trabalho tenha foco nos dispositivos móveis, os experimentos foram executados usando um computador pessoal, pois o ambiente WEKA não é disponibilizado para dispositivos móveis. Isso representa uma limitação e ameaça a validade dessa pesquisa, entretanto, essa limitação impacta exclusivamente as medidas de eficiência por causa da maior capacidade de processamento e memória principal dos computadores pessoais, quando comparados com os dispositivos móveis em geral. Assim, as medidas de eficiência coletadas nesta avaliação não devem ser consideradas em valores absolutos e nem usadas para comparação com outros resultados obtidos com experimentos executados em outras plataformas com capacidade diferente. No entanto, os resultados obtidos com as medidas de eficácia nesta avaliação não são afetados por essa limitação e podem ser usados sem restrições.

5.4 Análise

O ambiente WEKA também foi utilizado para a análise dos resultados do experimento. Com ele, é possível selecionar uma das medidas de desempenho suportadas e sumarizar da forma desejada os resultados individuais obtidos nos experimentos. Os resultados das medidas de eficácia e eficiência definidas na subseção 5.2.3 foram sumarizadas por cenário de avaliação (conjunto de dados, conjunto de *features* e algoritmo de aprendizado). Para a sumarização foram utilizadas a média aritmética e o desvio padrão dos resultados obtidos no cenário, considerando os experimentos realizados para todos os usuários pertencentes ao conjunto de dados utilizado. Assim, foi possível determinar o desempenho obtido em cada cenário avaliado e comparar a

diferença entre eles.

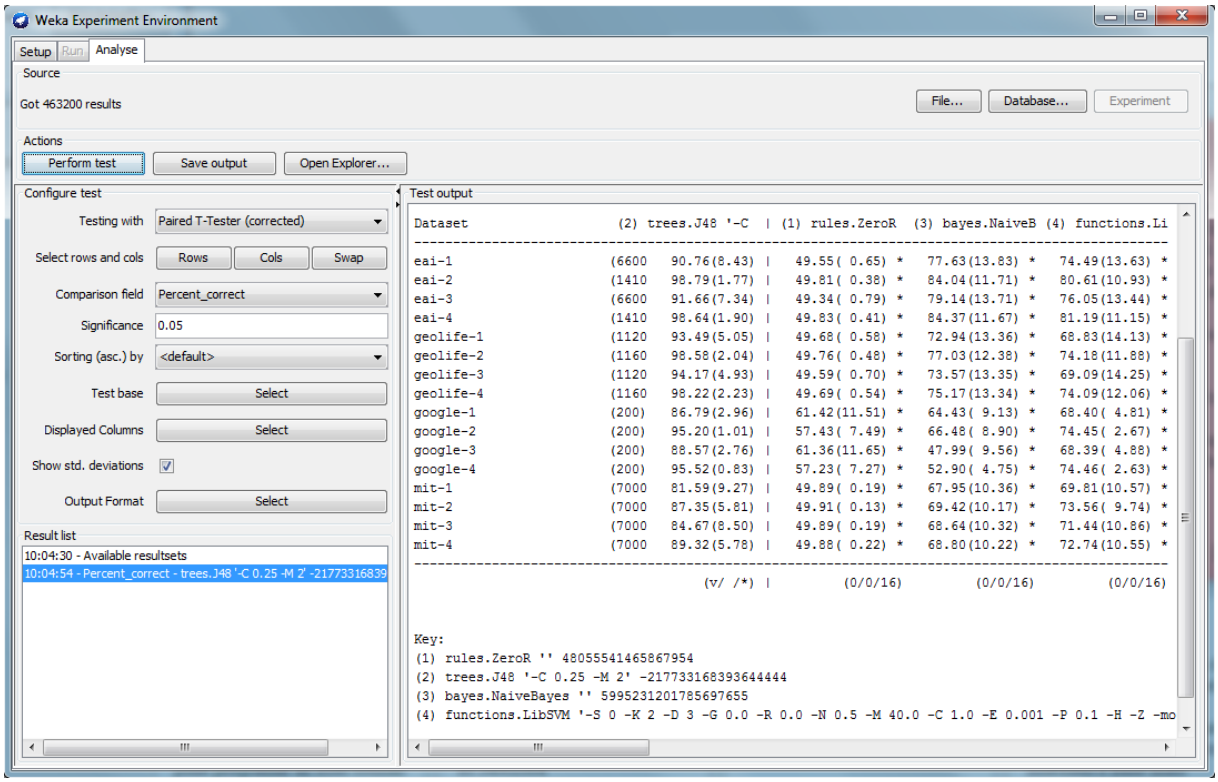
Para a comparação dos resultados sumarizados obtidos a partir de um mesmo conjunto de dados é necessário utilizar teste de hipótese. No entanto, do mesmo modo que as técnicas de repetições e validação cruzada discutido na seção 5.3, o teste de hipótese adequado para essa comparação representa uma questão investigada em alguns trabalhos na área de aprendizado de máquina (BOUCKAERT; FRANK, 2004)(DIETTERICH, 1998)(ALPAYDM, 1999), mas que permanece em aberto. Assim, nesta avaliação foi utilizado o Teste-T Pareado Corrigido (Paired Teste-T Corrected) , pois ele é suportado e sugerida por padrão pelo ambiente WEKA, tendo sido proposto pelos próprios autores da ferramenta (BOUCKAERT; FRANK, 2004). Além disso, segundo os resultados obtidos em (BOUCKAERT; FRANK, 2004), esse teste também aumenta a reprodutibilidade do experimento e isso está alinhado com os objetivos deste trabalho.

Outra comparação importante para a avaliação envolve os resultados obtidos a partir de diferentes conjuntos de dados. A média dos resultados obtidos nesses diferentes conjuntos de dados não pode ser usada para comparar dois algoritmos A e B (DEMŠAR, 2006). Nesse caso, a única informação que pode ser usada é o número de casos em que A tem desempenho melhor do que B, verificando a probabilidade disso ser por acaso e na verdade o desempenho dos dois ser igual (DEMŠAR, 2006). Os testes não paramétricos utilizam basicamente esse ranking e não os valores absolutos dos resultados. Assim, nesta avaliação foi utilizada a técnica de ranking suportada pelo ambiente WEKA. Ela contabiliza o número de casos em que um algoritmo foi melhor que os outros a partir de um mesmo conjunto de dados e utiliza essa informação para gerar uma classificação.

A Figura 10 apresenta a interface para análise de resultados de experimentos do ambiente WEKA. No caso, a interface está exibindo a análise dos resultados de todos os cenários avaliados. Na Figura 10 é possível visualizar as configurações utilizadas para análise dos resultados do experimento (*Configure test*). Por exemplo, o tipo de teste estatístico (*Testing with*), a medida de desempenho (*Comparison field*) e o nível de significância (*Significance*).

Além disso, na Figura 10 também é possível visualizar o resultado obtido na análise (*Test output*). Nesse resultado, as colunas representam os algoritmos de classificação avaliados e as linhas representam os outros fatores considerados nos cenários avaliados (conjuntos de *features* e conjuntos de dados). Assim, para cada cenário é possível obter a média e o desvio padrão da medida de comparação selecionada. Por exemplo, na Figura 10 é possível visualizar que o algoritmo DT (J48) obteve com o conjunto de dados Eai e o conjunto Features Set #1

Figura 10 – Interface para análise de resultados de experimento do WEKA



Fonte – O autor.

(eai-1) acurácia média 90,76% com desvio padrão 8,43.

Ainda nesse resultado é possível visualizar na primeira coluna dos algoritmos, que o classificador DT (J48) foi usado como base para o teste de hipótese. O resultado desse teste é representado por um símbolo ao lado do resultado do desempenho obtido pelos outros classificadores em cada cenário. O símbolo "*" indica que o resultado do classificador tomado como base para o teste foi significativamente maior do que o classificador da respectiva coluna, considerando o cenário avaliado na respectiva linha. Já o símbolo "v" indica que o resultado foi significativamente menor e a ausência desses símbolos indica que não houve diferença significativa entre os resultados. Por exemplo, na Figura 10 o resultado obtido pelo classificador DT (J48) foi significativamente maior do que todos os outros classificadores em todos os cenários avaliados.

5.5 Finalização

Nesta seção são apresentados os resultados das medidas obtidos nos experimentos. Primeiramente, os algoritmos são classificados de forma geral com base nas tabelas que apresentam os rankings gerados a partir dos resultados dos testes de hipótese estatístico aplicados.

Em seguida, os resultados das principais medidas são analisados de forma detalhada a partir dos gráficos de barra vertical que apresentam as médias dessas medidas e o respectivo intervalo de confiança bilateral (*two-sided confidence interval*) com nível de confiança 95%.

Como forma de demonstrar a aplicação do processo proposto e das recomendações seguidas neste trabalho (JAIN, 1990)(ALPAYDIN, 2014), a comparação dos resultados considerou todas as nove medidas de eficácia e as três medidas de eficiência coletadas nos experimentos conduzidos nesta avaliação. Em seguida, a análise detalhada considera apenas a acurácia e as medidas de eficiência (CPU utilizada para treinamento e teste; e memória utilizada para armazenamento). A acurácia foi escolhida entre as medidas de eficácia para a análise detalhada porque considera no seu cálculo as medida mais frequentemente utilizadas nos trabalhos relacionados (TP e FP), como apresentado na seção 3.2. Todas as medidas coletadas nos experimentos estão disponíveis para consulta e análise detalhada em um repositório no GitHub[4].

5.5.1 Eficácia

A Tabela 2 apresenta o ranking dos classificadores, considerando cada uma das medidas de eficácia selecionadas na subseção 5.2.3. Nessa tabela, os pontos representam a diferença entre o número de vezes que o classificador em questão obteve valor significativamente maior que os demais classificadores e o número de vezes que ele obteve valor significativamente menor, considerando a respectiva medida e todos os cenários avaliados nos experimentos. Essa diferença significativa foi avaliada usando a ferramenta de análise de experimentos do ambiente WEKA e o teste-T pareado corrigido com nível de significância de 0,05, conforme foi apresentado na seção 5.4. Assim, é possível identificar nesse ranking o classificador que obteve o melhor desempenho para cada uma das medidas de eficácia avaliadas. Embora o desempenho do algoritmo 0R não esteja apresentado na Tabela 2, ele foi usado como linha de base. É importante destacar que a medida FPR deve ser interpretada de forma inversa com relação às demais medidas. No caso da FPR, um valor menor representa um desempenho melhor do algoritmo.

Baseado nesse ranking, é possível perceber que o algoritmo DT obteve o melhor desempenho em todas as medidas de eficácia avaliadas. Na sequência, o algoritmo SVM obteve o segundo melhor desempenho em oito das nove medidas avaliadas (TPR, FPR, *Precision*, *Recall*, *Sensitivity*, *Specificity*, Acurácia e *F-Measure*). Por fim, o algoritmo NB obteve o segundo melhor desempenho em uma das nove medidas (AUC). Assim, as evidências indicam que o algoritmo

[4] https://github.com/GREatPesquisa/Seguranca/tree/master/Autenticacao_Transparente_Continua

DT obteve a maior eficácia na avaliação, seguido dos algoritmos SVM em segundo lugar e NB em terceiro.

Tabela 2 – Ranking dos algoritmos por medida de eficácia.

	DT			SVM			NB		
Medidas	Maior	Menor	**Pontos**	Maior	Menor	**Pontos**	Maior	Menor	**Pontos**
TPR	15	4	**11**	4	9	**-5***	3	10	**-7***
FPR	0	48	**-48**	17	13	**4**	18	10	**8**
Precision	48	0	**48**	19	16	**3**	14	21	**-7**
Recall	15	4	**11**	4	9	**-5***	3	10	**-7***
Sensitivity	15	4	**11**	4	9	**-5***	3	10	**-7***
Specificity	48	0	**48**	13	17	**-4**	10	18	**-8**
Acurácia	48	0	**48**	20	16	**4**	13	20	**-7**
F-Measure	46	0	**46**	13	15	**-2**	12	17	**-5**
AUC	47	0	**47**	16	28	**-12**	28	15	**13**

Fonte – O autor.

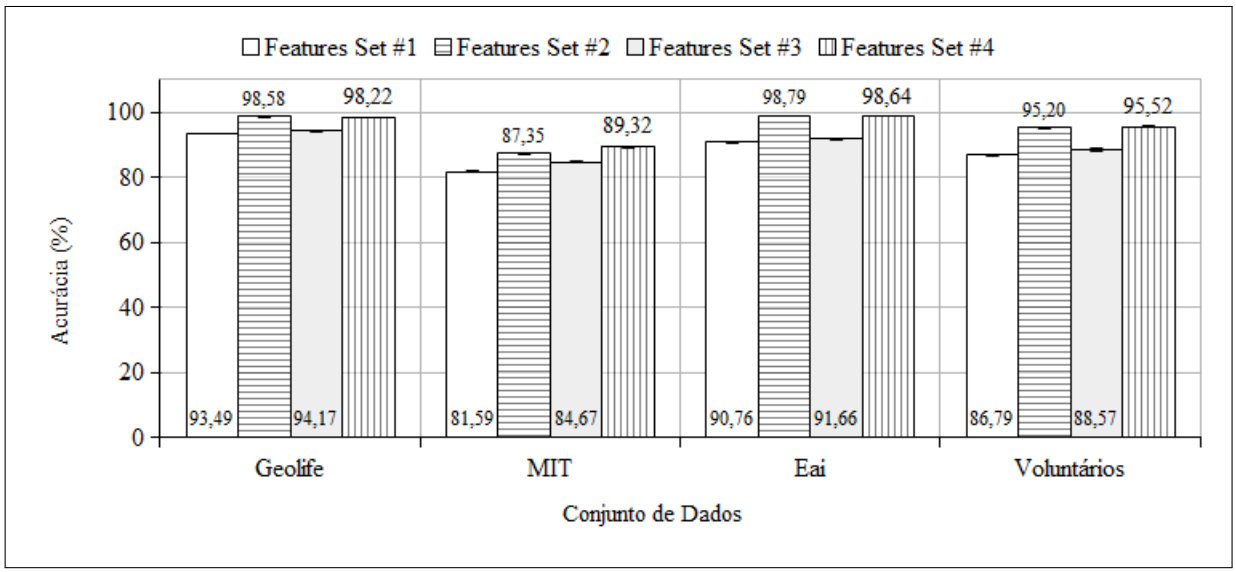
Nota – O símbolo "*" indica que o desempenho foi pior do que o algoritmo 0R, considerando o algoritmo e a medida em questão.

Acurácia. Iniciando a análise detalhada dos resultados das medidas de eficácia, as Figuras 11, 12 e 13 apresentam a média da acurácia obtida com os conjuntos de *features* avaliados, considerando respectivamente os algoritmos DT, SVM e NB. Essas médias sumarizam os resultados obtidos por conjunto de dados, considerando os resultados individuais de cada usuário do respectivo conjunto. Analisando essas figuras é possível perceber que o algoritmo DT apresentou acurácia de até 98,79%. Além disso, a acurácia apresentada pelo algoritmo DT representou um aumento de até 36,30% quando comparado com o segundo colocado (SVM), e de até 84,54% quando comparado com o terceiro colocado (NB).

Seguindo com a análise das Figuras 11, 12 e 13 é possível identificar que esses algoritmos apresentaram frequentemente acurácia maior usando os conjuntos Features Set #2 e Features Set #4. Esse aumento da acurácia é mais significativo e frequente nos resultados obtidos com o algoritmo DT. Esses conjuntos têm em comum a utilização do tempo juntamente com as *features* de localização. Assim, as evidências indicam que a utilização do tempo aumenta a acurácia das *features* localização.

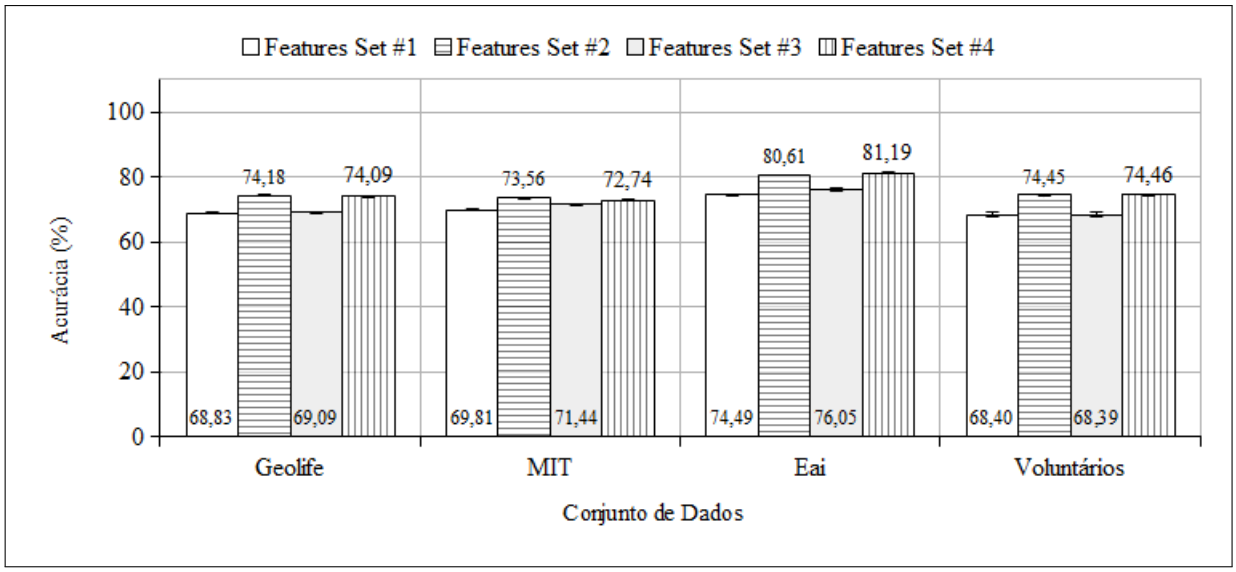
Continuando a análise das Figuras 11, 12 e 13 também é possível identificar que esses algoritmos não apresentaram aumento significativo da acurácia usando os conjuntos Features Set #3 e Features Set #4 quando comparados com a acurácia obtida com os conjuntos Features Set #1 e Features Set #2, respectivamente. Em alguns casos é possível observar o inverso, pois a utilização dos conjuntos Features Set #3 e Features Set #4 gerou a diminuição da acurácia. Isso

Figura 11 – Acurácia com o algoritmo DT



Fonte – O autor.

Figura 12 – Acurácia com o algoritmo SVM

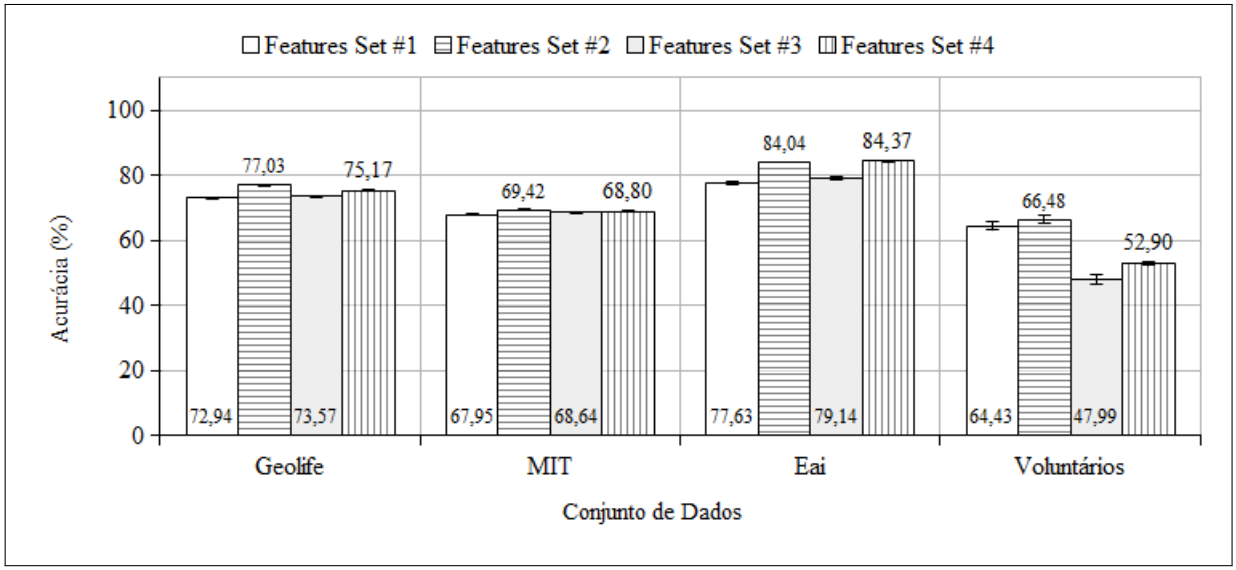


Fonte – O autor.

pode ser observado com mais frequência no caso do conjunto Features Set #4. Além disso, essa diminuição pode ser observada com maior intensidade na Figura 13, onde apresenta a acurácia do algoritmo NB com o conjunto Voluntários. Esses dois primeiros conjuntos de *features* têm em comum a utilização das sequências de localizações recentes do usuários como *features*. Assim, as evidências indicam que a utilização dessas sequência de localizações recentes não aumenta a acurácia e pode em alguns caso até mesmo provocar a sua diminuição.

Ainda analisando as Figuras 11, 12 e 13 é possível identificar que as *features* extraídas

Figura 13 – Acurácia com o algoritmo NB



Fonte – O autor.

do conjunto de dados MIT apresentaram frequentemente acurácia menor quando comparadas com o resultado obtido com o mesmo conjunto de *features* extraídas dos conjuntos Geolife e Eai. Essa diminuição da acurácia é mais significativa e frequente nos resultados obtidos com os algoritmos DT e NB. Nos cenários envolvendo esses algoritmos, as *features* baseadas em GPS apresentaram respectivamente acurácia 10,60% e 21,53% maior do que as *features* baseadas em Sistema de Telefonia Celular, considerando as maiores acurácias apresentadas com essas *features*. Conforme apresentado na subseção 5.2.4, a diferença do conjunto MIT é que as localizações foram coletadas usando o sistema de telefonia celular. Já nos outros conjuntos de dados utilizados nesta avaliação as localizações foram coletados usando GPS. Assim, as evidências indicam que a utilização de dados de localização coletados a partir de sistema de telefonia celular diminuem a acurácia.

Por fim, concluindo a análise das Figuras 11, 12 e 13 é possível perceber que o conjunto de dados Voluntários apresentou frequentemente acurácia menor quando comparado com os outros conjuntos de dados coletados via GPS, considerando os mesmos conjuntos de *features*. Essa diminuição foi mais significativa e frequente nos resultados obtidos com os algoritmos DT e NB. Conforme apresentado na subseção 5.2.4, a diferença do conjunto Voluntários é que os seus usuários são conhecidos e possuem hábitos de localização semelhantes. Assim, uma justificativa para essa diminuição da acurácia no conjunto Voluntários pode ser o nível de semelhança mais elevado, quando comparado com os outros conjuntos. Portanto, existem evidências que a acurácia obtida com os conjuntos Geolife e Eai são superestimados

devido à baixa similaridade dos hábitos de localização dos usuários desses conjuntos de dados.

5.5.2 Eficiência

A Tabela 3 possui as mesmas características da Tabela 2, descrita em detalhes na subseção 5.5.1. No entanto, A Tabela 3 apresenta o ranking dos classificadores, considerando cada uma das medidas de eficiência selecionadas na subseção 5.2.3. Assim, é possível identificar nesse ranking o classificador que obteve o melhor desempenho para cada uma das medidas de eficiência avaliadas. É importante destacar que todas essas medidas de eficiência devem ser interpretadas de forma inversa. No caso, valores menores para a utilização de CPU e memória representam um desempenho melhor do algoritmo.

Tabela 3 – Ranking dos algoritmos por medida de eficiência.

	DT			**SVM**			**NB**		
Medidas	Maior	Menor	**Pontos**	Maior	Menor	**Pontos**	Maior	Menor	**Pontos**
CPU - treinamento	22	11	**11**	33	0	**33**	5	21	**-16**
CPU - teste	0	11	**-11**	30	0	**30**	2	10	**-8**
Memória	32	16	**16**	48	0	**48**	16	32	**-16**

Fonte – O autor.

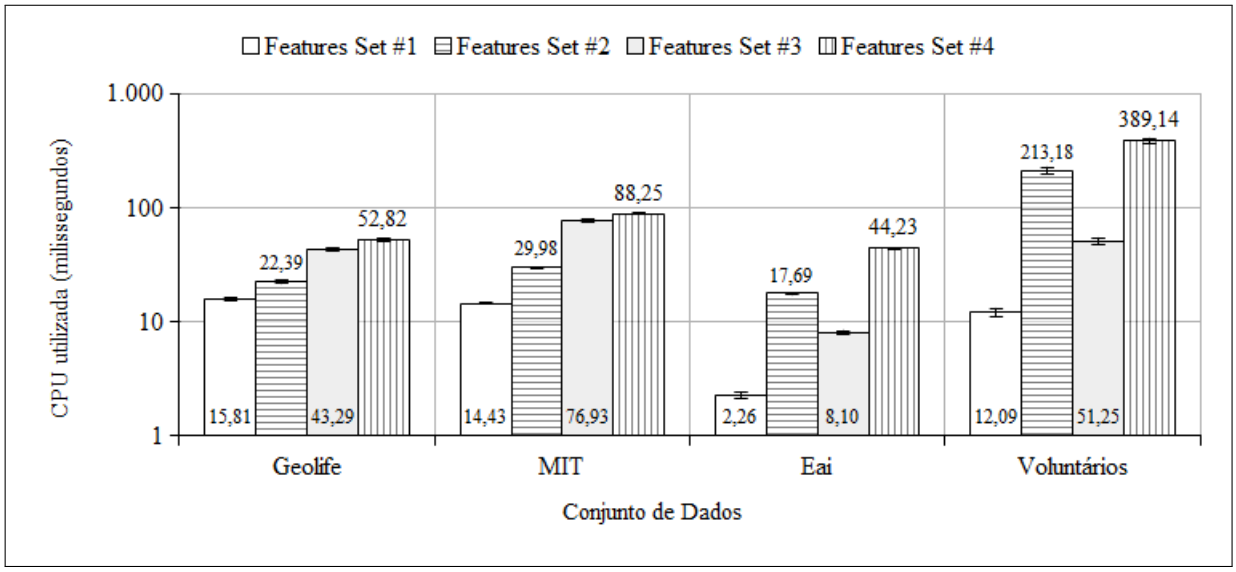
Baseado nesse ranking, é possível perceber que o algoritmo NB obteve o melhor desempenho em duas das três medidas de eficiência avaliadas (CPU - treinamento e Memória - armazenamento), seguido do algoritmo DT que ficou no segundo lugar nessas mesmas medidas. Na sequência, o algoritmo DT obteve o melhor desempenho em uma das três medidas de eficiência avaliadas (CPU - teste), seguido do algoritmo NB que ficou no segundo lugar nessa mesma medida. Por fim, o algoritmo SVM obteve o pior desempenho em todas as medidas de eficiência avaliadas. Assim, as evidências indicam que o algoritmo NB obteve a maior eficiência na avaliação, seguido dos algoritmos DT em segundo lugar e SVM em terceiro. No entanto, as evidências indicam também que o algoritmo DT apresenta maior eficiência que os outros algoritmos no processamento usando o modelo classificador já treinado. Essa característica é importante, pois é esperado que esse tipo de processamento seja demandado com maior frequência para a realização de predições do que o processamento para treinamento do modelo classificador nos mecanismos de autenticação transparente e contínua em dispositivos móveis.

CPU - Treinamento. Iniciando a análise detalhada dos resultados das medidas de eficiência, as Figuras 14, 15 e 16 apresentam a média do tempo de CPU utilizado para o

treinamento dos modelos classificadores gerados com os conjuntos de *features* avaliados, considerando respectivamente os algoritmos DT, SVM e NB. Como nas outras figuras apresentadas nesta seção, essas médias sumarizam os resultados obtidos por conjunto de dados, considerando os resultados individuais de cada usuário do respectivo conjunto. Analisando essas figuras é possível identificar que a redução do tempo de CPU utilizado para treinamento com o algoritmo NB foi de até 96,38% quando comparado com o segundo colocado (DT) e de até 99,92% quando comparado com o terceiro colocado (SVM).

Além disso, todos os algoritmos utilizaram mais tempo de CPU para treinamento com os conjuntos de *features* mais complexos (mais *features* por amostra). Nesse sentido, o conjunto Features Set #1 é o conjunto mais simples, incluindo apenas a última localização do usuário como feature. Já o conjunto Features Set #4 é o mais complexo, incluindo a sequência das últimas localizações do usuário e também o tempo da localização mais recente. Em alguns cenários (Eai e Voluntários) o aumento do tempo de CPU utilizado para treinamento foi mais significativo com os conjuntos Features Set #2 e Features Set #4, quando comparados respectivamente com os conjuntos Features Set #1 e Features Set #3. Os conjuntos Features Set #2 e Features Set #4 têm em comum a utilização do tempo juntamente com as *features* de localização. Assim, as evidências indicam que a utilização do tempo pode aumentar o tempo de CPU utilizado para treinamento.

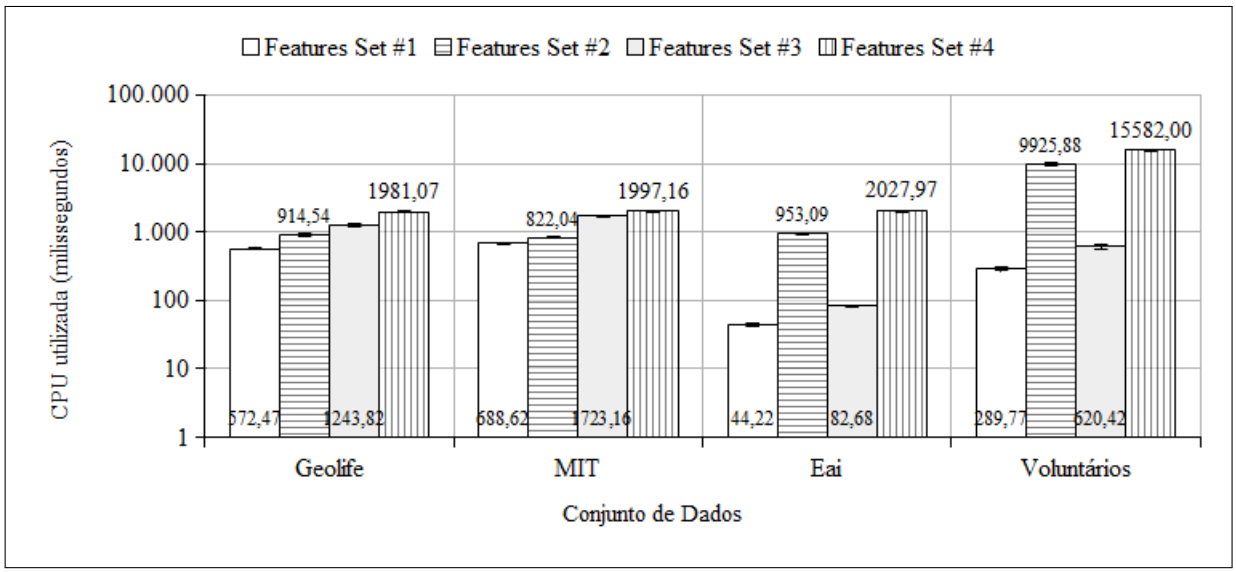
Figura 14 – CPU utilizada para treinamento com o algoritmo DT



Fonte – O autor.

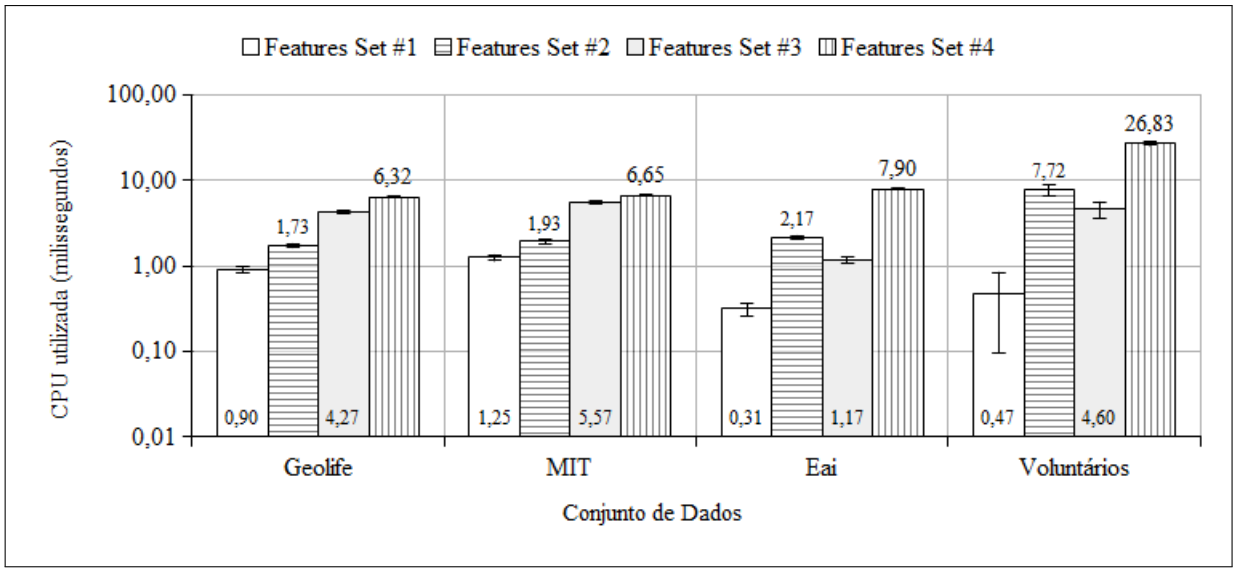
CPU - Teste. Na sequência, as Figuras 17, 18 e 19 apresentam a média do tempo de

Figura 15 – CPU utilizada para treinamento com o algoritmo SVM



Fonte – O autor.

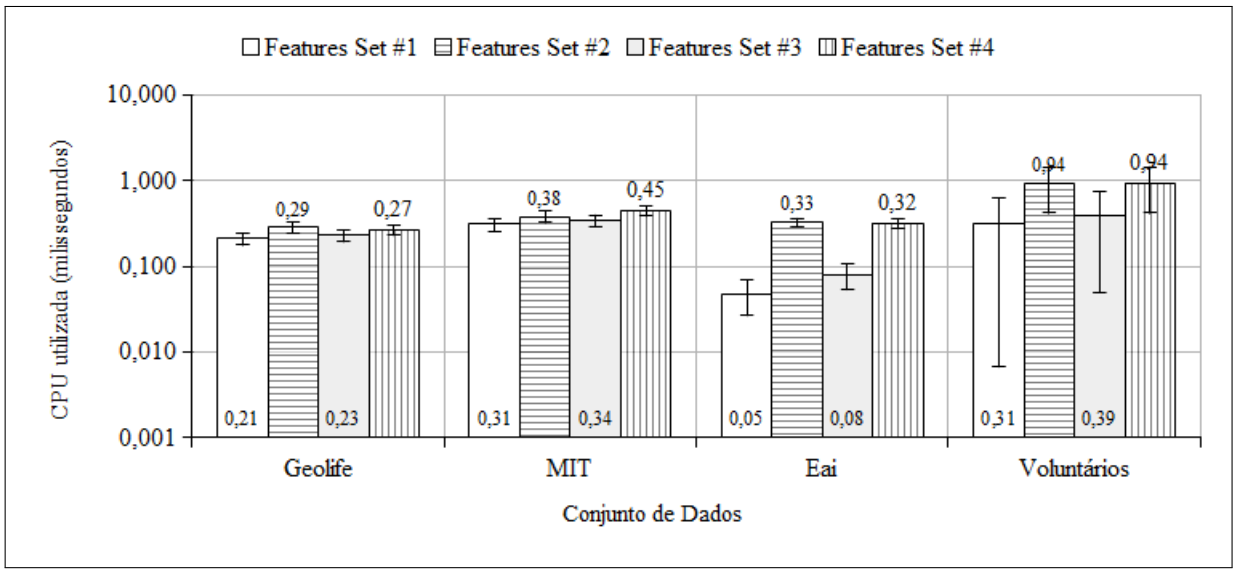
Figura 16 – CPU utilizada para treinamento com o algoritmo NB



Fonte – O autor.

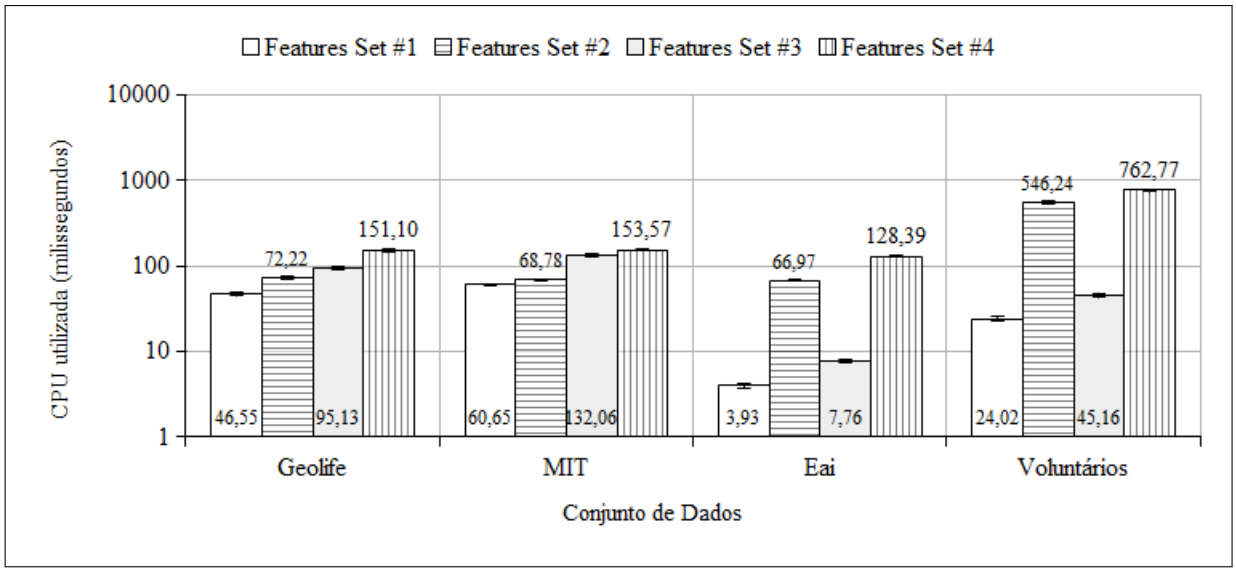
CPU utilizado para o teste dos modelos classificadores gerados com os conjuntos de *features* avaliados, considerando respectivamente os algoritmos DT, SVM e NB. Como nas outras figuras apresentadas nesta seção, essas médias sumarizam os resultados obtidos por conjunto de dados, considerando os resultados individuais de cada usuário do respectivo conjunto. Analisando essas figuras é possível identificar que a redução do tempo de CPU utilizado para teste com o algoritmo DT foi de até 90,00% quando comparado com o segundo colocado (NB) e de até 99,88% quando comparado com o terceiro colocado (SVM).

Figura 17 – CPU utilizada para teste com o algoritmo DT



Fonte – O autor.

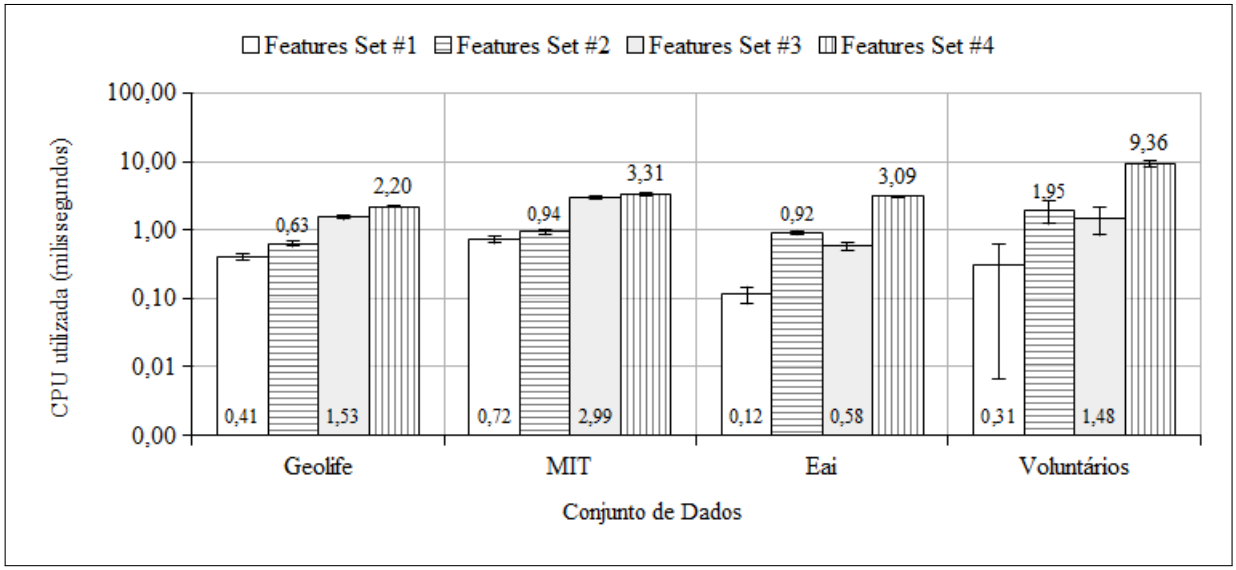
Figura 18 – CPU utilizada para teste com o algoritmo SVM



Fonte – O autor.

Além disso, o tempo de CPU utilizado para teste foi significativamente menor do que o tempo de CPU utilizado para treinamento em todos os cenários avaliados. Além disso, o algoritmo DT apresentou tempo de CPU utilizado para teste menor do que todos os outros algoritmos em todos os cenários. Ao mesmo tempo, o algoritmo DT apresentou em alguns cenários (Geolife e MIT) resultados mais constantes, independentemente do conjunto de *features* utilizado. Assim, as evidências indicam que o algoritmo DT é mais eficiente na utilização de tempo de CPU para predições e esse tempo para esse algoritmo pode ser menos dependente do

Figura 19 – CPU utilizada para teste com o algoritmo NB



Fonte – O autor.

conjunto de *features* de localização utilizado para a predição.

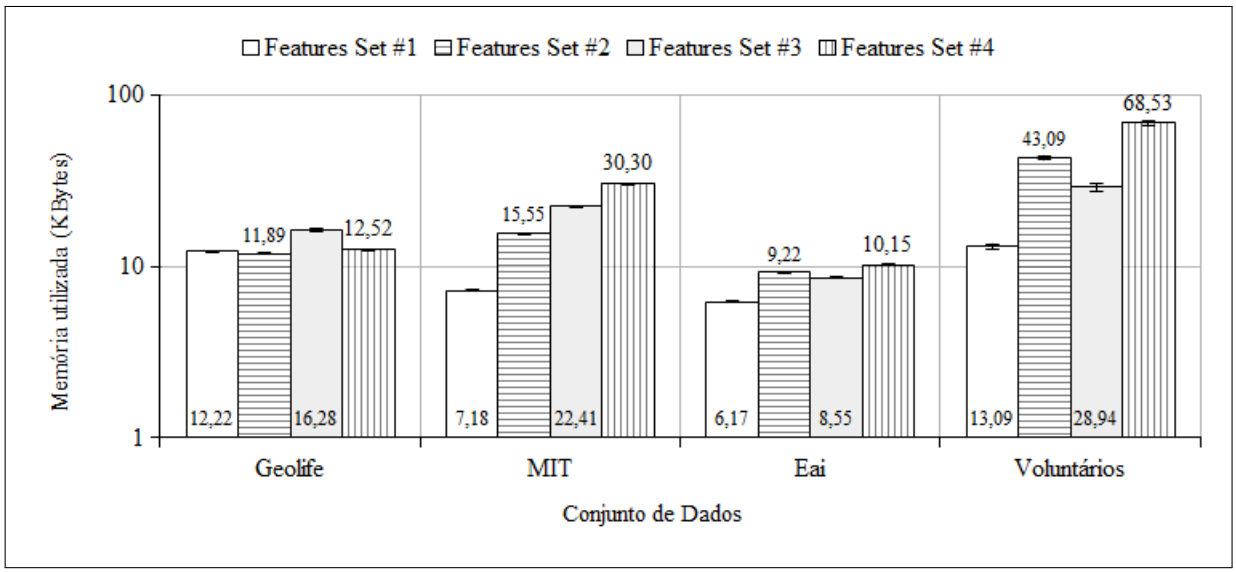
Memória. Por fim, as Figuras 20, 21 e 22 apresentam a média da memória utilizada para o armazenamento dos modelos classificadores gerados com os conjuntos de *features* avaliados, considerando respectivamente os algoritmos DT, SVM e NB. Como nas outras figuras apresentadas nesta seção, essas médias sumarizam os resultados obtidos por conjunto de dados, considerando os resultados individuais de cada usuário do respectivo conjunto. Analisando essas figuras é possível identificar que a redução da memória utilizada para armazenamento com o algoritmo NB foi de até 94,45%, quando comparado com o segundo colocado (DT), e de até a 99,73%, quando comparado com o terceiro colocado (SVM).

5.5.3 Resultado da Avaliação

Comparando os resultados dos experimentos conduzidos para a avaliação das *features* de localização usadas no problema de autenticação transparente e contínua foram obtidas evidências que o algoritmo DT oferece a maior eficácia, seguido respectivamente dos algoritmos SVM e NB em segundo e terceiro lugar. No entanto, as evidências também indicaram que o algoritmo NB oferece a maior eficiência, seguido dos algoritmos DT e SVM em segundo e terceiro lugar, respectivamente.

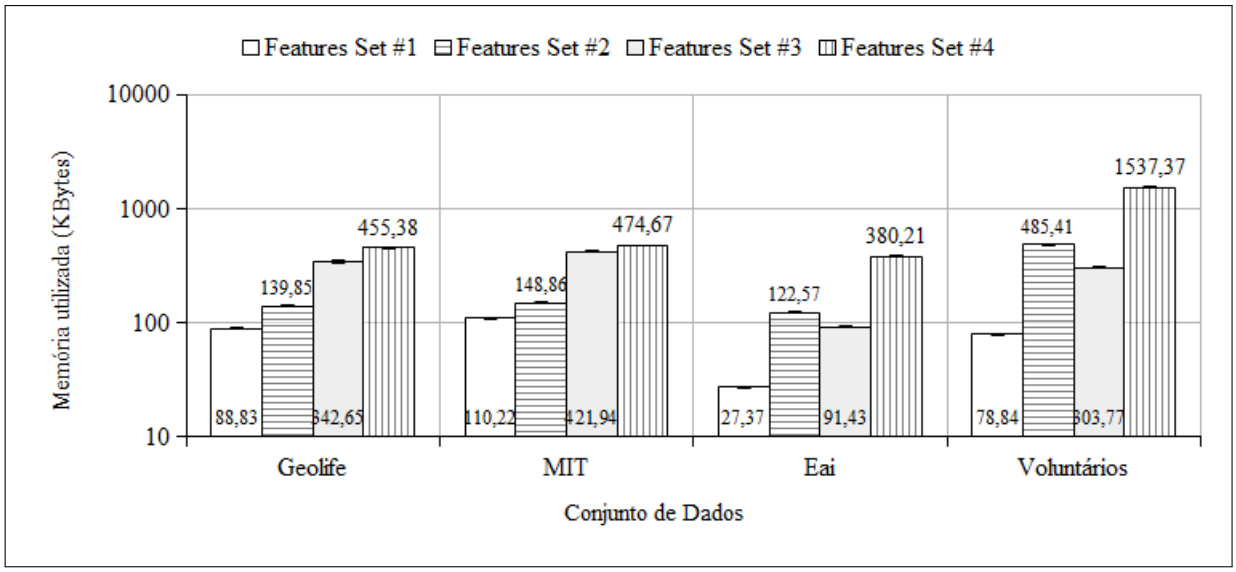
Além disso, a análise detalhada dos resultados também indicou que as *features* baseadas em GPS apresentam maior acurácia do que as *features* baseadas em Sistema de

Figura 20 – Memória utilizada para armazenamento com o algoritmo DT



Fonte – O autor.

Figura 21 – Memória utilizada para armazenamento com o algoritmo SVM

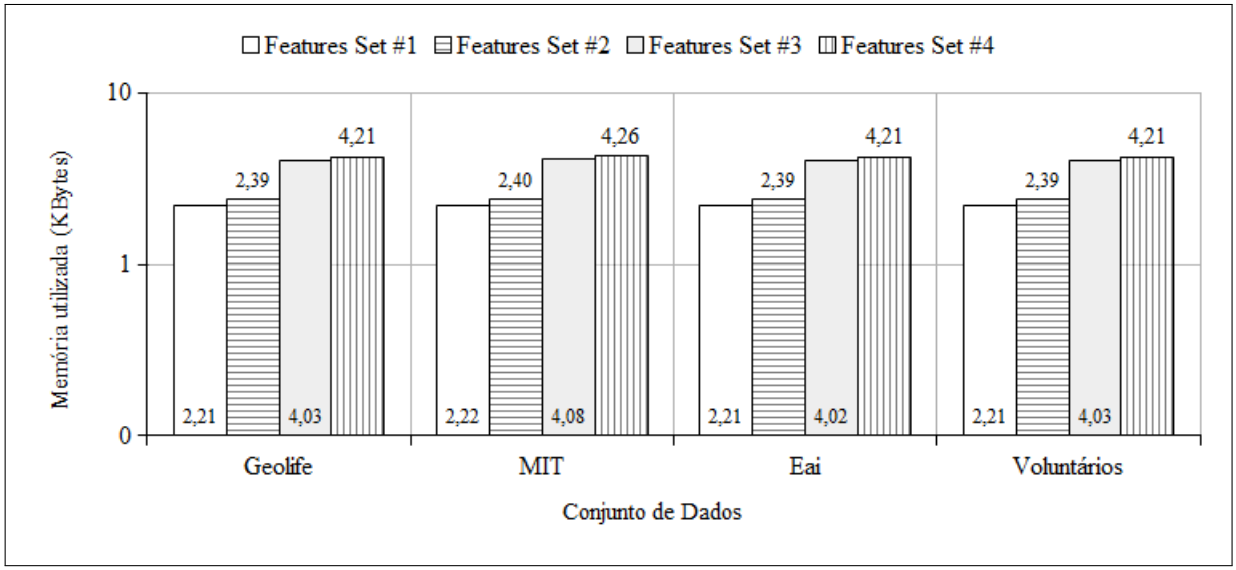


Fonte – O autor.

Telefonia Celular. Além disso, as evidências também indicam que o uso de *features* de localização mescladas com *features* de tempo (Features Set #2 e Features Set #4) também aumentam a acurácia. No entanto, o uso de *features* que mesclam a sequência das últimas localizações e o tempo (Features Set #4) diminuem a acurácia e a eficiência, quando comparadas com as *features* que mesclam apenas a última localização com o tempo (Features Set #2).

Por fim, a acurácia obtida com os experimentos usando o conjunto Voluntários pode representar evidência que as medidas de eficácia obtidas com os conjuntos Geolife e Eai são

Figura 22 – Memória utilizada para armazenamento com o algoritmo NB



Fonte – O autor.

superestimadas. Um baixo nível de semelhança entre os usuários desse conjunto de dados pode não representar as ameaças reais a que um mecanismo de autenticação está exposto e superestimar o resultado das medidas de eficácia avaliadas neste trabalho.

5.6 Considerações Finais

Neste capítulo foram apresentados os detalhes da avaliação das *features* de localização identificadas nos trabalhos relacionados apresentados na seção 3.1. Essa avaliação seguiu as atividades de planejamento, execução, análise e finalização previstas no processo proposto e apresentadas no Capítulo 4.

Assim, os resultados apresentados neste capítulo produzem evidência da viabilidade da utilização desse processo de avaliação de *features* biométricas para autenticação transparente e contínua em dispositivos móveis usando aprendizado de máquina. Além disso, eles também podem ser utilizados para subsidiar a escolha do conjunto de *features* de localização e do algoritmo de aprendizado de máquina na construção de novos mecanismos de autenticação transparente e contínuos. Esses resultados podem ser utilizados também na comparação do desempenho de novas propostas de *features* de localização para autenticação transparente e contínua.

6 CONCLUSÃO

Este capítulo está estruturado da seguinte forma: a seção 6.1 lista os resultados alcançados e os artigos produzidos durante a realização da pesquisa; na sequência, a seção 6.2 discute as limitações deste trabalho; e, por fim, a seção 6.3 apresenta as oportunidades de trabalhos futuros.

6.1 Resultados Alcançados

Com os dispositivos móveis nasceu também uma nova dinâmica de utilização dos usuários, gerando a necessidade de repensar a abordagem de autenticação comumente utilizados nos sistemas (JAKOBSSON *et al.*, 2009)(CRAWFORD, 2014). Na busca por atender essa necessidade surgiram várias abordagens de autenticação, dentre elas a autenticação transparente e contínua (CLARKE, 2011) que busca construir sistemas biométricos baseados no reconhecimento de padrões.

Neste trabalho, assim como em (KHAN, 2016) e (MARCEL, 2013), foram identificadas algumas deficiências nas avaliações de propostas de mecanismos de autenticação baseados nessa nova abordagem. Essas deficiências dificultam ou até mesmo impedem a comparação do desempenho desses mecanismos, a reprodução dos resultados apresentados nesses trabalhos e, consequentemente, o reuso das técnicas utilizadas em novas pesquisas para construção de sistemas com melhor desempenho.

Adicionalmente, este trabalho também identificou semelhanças entre esses mecanismos de autenticação e os sistemas de aprendizado de máquina, inclusive nas questões relacionadas à avaliação. Em ambos, as *features* representam uma das principais técnicas utilizadas, senão a principal, com influência direta e significativa sobre o desempenho. Outra técnica que também apresenta forte influência sobre o desempenho nesses dois tipos de sistemas computacionais é o algoritmo de aprendizado utilizado. No entanto, já existe um amplo conjunto de algoritmos de aprendizado de máquina de propósito geral bem conhecidos e estudados, que podem ser utilizados nesses sistemas. Já as *features* biométricas são específicas de uma proposta e comumente poucos detalhes sobre elas são apresentados.

Com base nesse contexto, este trabalho reuniu um conjunto de recomendações de boas práticas nas áreas de avaliação de desempenho de sistemas computacionais, avaliação de sistemas biométricos e experimentos de aprendizado de máquina. A partir dessas recomendações,

este trabalho propôs um processo de avaliação baseado em experimentos de aprendizado de máquina com foco no desempenho das *features* biométricas utilizadas nos mecanismos de autenticação transparente e contínua em dispositivos móveis.

O processo de avaliação proposto neste trabalho inclui nove atividades: Definir Objetivo; Modelar Problema; Definir Variáveis de Resposta; Definir Fatores e Níveis; Executar Pré-testes; Definir Design do Experimento; Executar Experimentos; Analisar Resultados; e Apresentar Resultados. Essas atividades foram divididas em quatros especialidades: Planejamento; Execução; Análise; e Finalização. Esse processo representa um primeiro passo no sentido de padronizar a avaliação e comparação do desempenho desses mecanismos de autenticação. Com isso, busca-se melhorar a reprodutibilidade e a granularidade dessas avaliações. Além disso, busca-se também potencializar a reutilização das *features* biométricas utilizadas na autenticação transparente e contínua.

Para aplicar na prática o processo proposto, este trabalho realizou um estudo de caso com uma avaliação das *features* de localização *outdoor* utilizadas nos trabalhos relacionados identificados durante a revisão da literatura. Esse estudo de caso produziu evidências sobre a viabilidade da utilização do processo proposto para a avaliação do desempenho de *features* biométricas usadas para a autenticação transparente e contínua em dispositivos móveis. Essa avaliação gerou também resultados que depois de analisados permitiram comparar o desempenho das técnicas utilizadas na literatura e identificar o conjunto de *features* de localização que apresenta melhor eficácia e eficiência. Além disso, permitiu também identificar a diferença de desempenho entre as técnicas de localização *outdoor* (GPS e Sistema de Telefonia Celular) e os algoritmos de aprendizado utilizadas nesses trabalhos. Com esses resultados, essas técnicas que apresentaram melhor desempenho podem ser reutilizadas por outros pesquisadores, combinadas com *features* obtidas de outros traços biométricos para a construção de novas propostas de mecanismos de autenticação transparente e contínua.

Por último, é importante apresentar os artigos científicos produzidos durante o período de mestrado:

- **CORREIA, M. A. S.**; XIMENES, P.; ANDRADE, R. M. C. Sistematização do Contexto como Fator de Autenticação de Usuários de Dispositivos Móveis. Resumo Estendido. In: Anais do XV Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg 2015), v. 1, p. 358-361. SBC, Porto Alegre, 2015; e

- XIMENES, P.; **CORREIA, M. A. S.**; MELLO, P.; CARVALHO, F.; FRANKLIN, M.; ANDRADE, R. M. C. TARP Fingerprinting: Um Mecanismo de Browser Fingerprinting Baseado em HTML5 Resistente a Contramedidas. Artigo Completo. In: Anais do XVI Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg 2016), v. 1, p. 100-113. SBC, Porto Alegre, 2015.

6.2 Limitações

Os experimentos conduzidos nesta pesquisa para a avaliação das *features* de localização *outdoor* não foram executados em dispositivos móveis, pois não foi encontrado um ambiente de execução de experimentos de aprendizado de máquina que atendesse esse requisito e estava fora do escopo deste trabalho o desenvolvimento de um ambiente próprio com essa característica. Isso representa uma limitação para esse trabalho, pois os resultados das medidas de eficiência relacionadas ao tempo de CPU utilizado e mensurado durante a realização do estudo de caso não podem ser analisados em valores absolutos, porque a capacidade de processamento do computador utilizado para a execução dos experimentos é superior à de um dispositivo móvel convencional. No entanto, buscando minimizar esta ameaça à validade desta pesquisa, todos os experimentos foram executados no mesmo ambiente e os resultados especificamente dessa medida para os diferentes conjuntos de *features* avaliados foram analisados apenas de forma relativa.

Outra questão que pode ser considerada uma limitação deste trabalho diz respeito à avaliação do processo proposto. Embora este trabalho considere que não existam particularidades que impeçam a avaliação de outras *features* biométricas usando esse mesmo processo, a avaliação realizada utilizou apenas as *features* de localização *outdoor*. Além disso, a avaliação realizada neste trabalho foi conduzida pela mesmo pesquisador responsável pela elaboração do processo proposto. Isso representa uma ameaça à validade desta pesquisa e ela foi mitigada buscando fornecer ao máximo os meios para que os resultados obtidos neste trabalho possam ser reproduzidos por terceiros. Para isso, foram utilizadas ferramentas livre e dados públicos para a realização do estudo de caso, e também foram descritos todos os detalhes e decisões tomadas na avaliação.

6.3 Trabalhos Futuros

Neste trabalho foi possível perceber a necessidade de avanço na avaliação das propostas de autenticação transparente e contínua para dispositivos móveis, principalmente no que diz respeito a medidas, processos e ferramentas. Esse avanço deve buscar a padronização das avaliações realizadas pelos trabalhos nesta área de pesquisa, garantindo assim a comparação entre os resultados obtidos com as diferentes técnicas propostas pelos autores. Além disso, esses avanços também devem buscar a reprodutibilidade dos experimentos, promovendo assim a validação por terceiros dos resultados apresentados e o reuso das técnicas propostas. Esta dissertação contribuiu com uma proposta de processo com foco na avaliação das *features* biométricas utilizadas. Assim, ainda representam oportunidades de trabalhos futuros:

- **Padronização de medidas.** Embora a autenticação transparente e contínua utilize técnicas provenientes dos sistemas biométricos e existam na literatura várias medidas sugeridas para a avaliação desses sistemas, ainda é possível perceber nessas duas áreas a falta de padronização entre as medidas utilizadas para a avaliação dos trabalhos propostos (MARCEL, 2013);
- **Padronização de outros processos.** Apesar deste trabalho propor um processo para a avaliação de desempenho de *features* biométricas, outros aspectos como a usabilidade (JAIN, 1990), a privacidade e as vulnerabilidades (MARCEL, 2013) também precisam ser avaliados nesses sistemas. Da mesma forma que o desempenho, para que os resultados das avaliações desses outros aspectos também sejam comparáveis, reproduzíveis e reutilizáveis é necessária a utilização de processos padronizados; e
- **Padronização de ferramentas e dados para realização dos experimentos.** Foram identificados na literatura alguns esforços no sentido de padronização de ferramentas e dados para a avaliação biométrica (MARCEL, 2013)(ANJOS *et al.*, 2012). No entanto, essas ferramentas não são utilizadas pelos trabalhos relacionados encontrados na literatura durante esta pesquisa . Além disso, elas fornecem suporte apenas a um conjunto limitado de traços biométricos.

Além disso, com base nas limitações desta dissertação, também é possível identificar algumas oportunidades de trabalhos futuros. São elas:

- **Avaliação do processo proposto.** Para avaliar a viabilidade do processo pro-

posto neste trabalho é necessária o sua utilização por outros pesquisadores na avaliação de *features* extraídas de traços biométricos, preferencialmente traços biométricos diferentes da localização *outdoor* usada neste trabalho. Como resultado desse trabalho futuro identificado, devem ser produzidas novas evidências sobre a viabilidade do processo proposto ou a necessidade de ajustes. Além disso, devem ser produzidos também novos estudos de caso que possam ser utilizados por outros pesquisadores na criação de novas propostas de mecanismos de autenticação transparente e contínua; e

- **Desenvolvimento de ferramenta de avaliação para dispositivos móveis.** Para avaliar melhor o desempenho das propostas de *features* biométricas para autenticação transparente e contínua em dispositivos móveis usando aprendizado de máquina é necessário o desenvolvimento de uma ferramenta com suporte a plataforma móvel. A necessidade dessa ferramenta é ainda maior no caso da avaliação com foco nas medidas de utilização de recursos computacionais.

REFERÊNCIAS

ALPAYDIN, E. **Introduction to machine learning**. [S.l.]: MIT press, 2014.

ALPAYDM, E. Combined 5× 2 cv f test for comparing supervised classification learning algorithms. **Neural computation**, MIT Press, v. 11, n. 8, p. 1885–1892, 1999.

ANJOS, A.; EL-SHAFEY, L.; WALLACE, R.; GÜNTHER, M.; MCCOOL, C.; MARCEL, S. Bob: a free signal processing and machine learning toolbox for researchers. In: ACM. **Proceedings of the 20th ACM international conference on Multimedia**. [S.l.], 2012. p. 1449–1452.

BOUCKAERT, R. R.; FRANK, E. Evaluating the replicability of significance tests for comparing learning algorithms. In: SPRINGER. **Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining**. [S.l.], 2004. p. 3–12.

BUTHPITIYA, S.; DEY, A. K.; GRISS, M. Soft authentication with low-cost signatures. In: IEEE. **Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on**. [S.l.], 2014. p. 172–180.

CAVALCANTE, T. M. **Avaliação de desempenho de mecanismos de segurança utilizados para prover os serviços de confidencialidade, integridade e autenticação em redes de sensores sem fio**. Dissertação (Mestrado) — Universidade Federal do Ceará, 2012.

CBOK, B. Guia para o gerenciamento de processos de negócio corpo comum de conhecimento. **Association of Business Process Management Professionals. ABPMP BPM CBOK**, 2013.

CHANG, C.-C.; LIN, C.-J. Libsvm: a library for support vector machines. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM, v. 2, n. 3, p. 27, 2011.

CHOW, R.; JAKOBSSON, M.; MASUOKA, R.; MOLINA, J.; NIU, Y.; SHI, E.; SONG, Z. Authentication in the clouds: a framework and its application to mobile users. In: ACM. **Proceedings of the 2010 ACM workshop on Cloud computing security workshop**. [S.l.], 2010. p. 1–6.

CLARKE, N. **Transparent user authentication: biometrics, RFID and behavioural profiling**. [S.l.]: Springer Science & Business Media, 2011.

CORMEN, T. H. **Introduction to algorithms**. [S.l.]: MIT press, 2009.

CRAWFORD, H. Adventures in authentication–position paper. In: USENIX ASSOCIATION. **Proceedings of the 2014 Symposium on Usable Privacy and Security (SOUPS)**. [S.l.], 2014.

DARIN, T.; ANDRADE, R. M. C.; MACEDO, J. A. F.; ARAÚJO, D.; MESQUITA, L.; SÁNCHEZ, J. Usability and ux evaluation of a mobile social application to increase students-faculty interactions. In: SPRINGER. **International Conference on Human-Computer Interaction**. [S.l.], 2016. p. 21–29.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine learning research**, v. 7, n. Jan, p. 1–30, 2006.

DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. **Neural computation**, MIT Press, v. 10, n. 7, p. 1895–1923, 1998.

DOMINGOS, P. A few useful things to know about machine learning. **Communications of the ACM**, ACM, v. 55, n. 10, p. 78–87, 2012.

EAGLE, N.; PENTLAND, A. S. Reality mining: sensing complex social systems. **Personal and ubiquitous computing**, Springer, v. 10, n. 4, p. 255–268, 2006.

FRIDMAN, L.; WEBER, S.; GREENSTADT, R.; KAM, M. Active authentication on mobile devices via stylometry, application usage, web browsing, and gps location. **IEEE Systems Journal**, IEEE, 2016.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM, v. 11, n. 1, p. 10–18, 2009.

HAYASHI, E.; DAS, S.; AMINI, S.; HONG, J.; OAKLEY, I. Casa: context-aware scalable authentication. In: ACM. **Proceedings of the Ninth Symposium on Usable Privacy and Security**. [S.l.], 2013.

HSU, C.-W.; CHANG, C.-C.; LIN, C.-J. *et al.* **A practical guide to support vector classification**. [S.l.], 2003. Disponível em: <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>. Acesso em: 2016-08-30.

ISO/IEC-15408-1: Information technology - security techniques - evaluation criteria for it security : Part 1 : introduction and general model. [S.l.]: ISO, 2009.

JAIN, A.; FLYNN, P.; ROSS, A. A. **Handbook of biometrics**. [S.l.]: Springer Science & Business Media, 2007.

JAIN, R. **The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling**. [S.l.]: John Wiley & Sons, 1990.

JAKOBSSON, M.; SHI, E.; GOLLE, P.; CHOW, R. Implicit authentication for mobile devices. In: USENIX ASSOCIATION. **Proceedings of the 4th USENIX conference on Hot topics in security**. [S.l.], 2009. p. 1–9.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: MORGAN KAUFMANN PUBLISHERS INC. **Proceedings of the Eleventh conference on Uncertainty in artificial intelligence**. [S.l.], 1995. p. 338–345.

JUST, M. Authentication frequency as an important design factor. In: USENIX ASSOCIATION. **Proceedings of the 2014 Symposium on Usable Privacy and Security (SOUPS)**. [S.l.], 2014.

KAYACIK, H. G.; JUST, M.; BAILLIE, L.; ASPINALL, D.; MICALLEF, N. Data driven authentication: On the effectiveness of user behaviour modelling with mobile device sensors. In: IEEE. **Proceedings of the Third Workshop on Mobile Security Technologies (MoST)**. [S.l.], 2014.

KHAN, H. **Evaluating the Efficacy of Implicit Authentication Under Realistic Operating Scenarios**. Tese (Doutorado) — University of Waterloo, 2016.

KHAN, H.; ATWATER, A.; HENGARTNER, U. Itus: an implicit authentication framework for android. In: ACM. **Proceedings of the 20th annual international conference on Mobile computing and networking**. [S.l.], 2014. p. 507–518.

KITCHENHAM, B.; CHARTERS, S. **Guidelines for performing systematic literature reviews in software engineering**. [S.l.], 2007. Disponível em: <https://www.cs.auckland.ac.nz/~mria007/Sulayman/Systematic_reviews_5_8.pdf>. Acesso em: 2016-08-30.

LIMA, J. C. D.; ROCHA, C. C.; VIEIRA, M. A.; AUGUSTIN, I.; DANTAS, M. A. Cars-ad: a context-aware recommender system to decide about implicit or explicit authentication in ubihealth. In: ACM. **Proceedings of the 9th ACM international symposium on Mobility management and wireless access**. [S.l.], 2011. p. 83–92.

LOOKOUT; SPRINT. **Sprint and Lookout consumer mobile behavior survey**. 2014. Disponível em: <http://blog.lookout.com/blog/2013/10/21/sprint-and-lookout-survey/>. Acesso em: 2016-08-30.

MAIA, M. E.; FONTELES, A.; NETO, B.; GADELHA, R.; VIANA, W.; ANDRADE, R. M. C. Loccam-loosely coupled context acquisition middleware. In: ACM. **Proceedings of the 28th Annual ACM Symposium on Applied Computing**. [S.l.], 2013. p. 534–541.

MARCEL, S. BEAT–biometrics evaluation and testing. **Biometric technology today**, Elsevier, v. 2013, n. 1, p. 5–7, 2013.

MELICHER, W.; KURILOVA, D.; SEGRETI, S. M.; KALVANI, P.; SHAY, R.; UR, B.; BAUER, L.; CHRISTIN, N.; CRANOR, L. F.; MAZUREK, M. L. Usability and security of text passwords on mobile devices. In: ACM. **Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems**. [S.l.], 2016. p. 527–539.

MICHAELIS: Dicionário brasileiro da língua portuguesa. Editora Melhoramentos, 2015. Disponível em: <http://michaelis.uol.com.br/moderno-portugues>. Acesso em: 2016-08-30.

MITCHELL, T. M. **Machine learning**. [S.l.]: McGraw-Hill, 1997.

MONTGOMERY, D. C. **Design and Analysis of Experiments**. 8. ed. [S.l.]: John Wiley & Sons, 2012.

NAKAMURA, E. T.; GEUS, P. L. de. **Segurança de redes em ambientes cooperativos**. [S.l.]: Novatec Editora, 2007.

PATEL, V. M.; CHELLAPPA, R.; CHANDRA, D.; BARBELLO, B. Continuous user authentication on mobile devices: Recent progress and remaining challenges. **IEEE Signal Processing Magazine**, IEEE, v. 33, n. 4, p. 49–61, 2016.

PHILLIPS, P. J.; MARTIN, A.; WILSON, C. L.; PRZYBOCKI, M. An introduction evaluating biometric systems. **Computer**, IEEE, v. 33, n. 2, p. 56–63, 2000.

QUINLAN, J. R. **C4. 5: programs for machine learning**. [S.l.]: Elsevier, 2014.

RAMAKRISHNAN, A.; TOMBAL, J.; PREUVENEERS, D.; BERBERS, Y. Prism: Policy-driven risk-based implicit locking for improving the security of mobile end-user devices. In: ACM. **Proceedings of the 13th International Conference on Advances in Mobile Computing and Multimedia**. [S.l.], 2015. p. 365–374.

RENAUD, K. Evaluating authentication mechanisms. In: CRANOR, L. F.; GARFINKEL, S. (Ed.). **Security and usability: design secure systems that people can use**. [S.l.]: O'Reilly, 2005. cap. 6, p. 103–128.

ROCHA, C. L.; BRILHANTE, I. R.; LETTICH, F.; MACEDO, J. A. F.; RAFFAETÀ, A.; ANDRADE, R. M. C.; ORLANDO, S. Tpred: a spatio-temporal location predictor framework. In: ACM. **Proceedings of the 20th International Database Engineering & Applications Symposium**. [S.l.], 2016. p. 34–42.

SHI, W.; YANG, J.; JIANG, Y.; YANG, F.; XIONG, Y. Senguard: Passive user identification on smartphones using multiple sensors. In: IEEE. **2011 IEEE 7th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)**. [S.l.], 2011. p. 141–148.

TANG, Y.; HIDENORI, N.; URANO, Y. User authentication on smart phones using a data mining method. In: IEEE. **2010 International Conference on Information Society (i-Society)**. [S.l.], 2010. p. 173–178.

THORNTON, C.; HUTTER, F.; HOOS, H. H.; LEYTON-BROWN, K. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In: ACM. **Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2013. p. 847–855.

VANDEWALLE, P.; KOVACEVIC, J.; VETTERLI, M. Reproducible research in signal processing. **IEEE Signal Processing Magazine**, IEEE, v. 26, n. 3, 2009.

ZHENG, Y.; XIE, X.; MA, W.-Y. Geolife: A collaborative social networking service among user, location and trajectory. **IEEE Data Eng. Bull.**, Citeseer, v. 33, n. 2, p. 32–39, 2010.

ZHOU, C.; FRANKOWSKI, D.; LUDFORD, P.; SHEKHAR, S.; TERVEEN, L. Discovering personally meaningful places: An interactive clustering approach. **ACM Transactions on Information Systems (TOIS)**, ACM, v. 25, n. 3, p. 12, 2007.